IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:
Florian v. Mulbe et al.

Application No.: 10/729,830 | Confirmation No.: 8653

Filed: December 5, 2003 | Art Unit: 1636

For: PHARMACEUTICAL COMPOSITION CONTAINING A STABILISED mRNA OPTIMISED FOR TRANSLATION IN ITS CODING REGIONS | Examiner: J. A. Dunston

THIRD DECLARATION OF DR. INGMAR HOERR

Dr. Ingmar Hoerr deposes and states as follows:

1. I am one of the inventors named on the above-captioned patent application, and I am affiliated with the assignee of this application. I provided a first declaration in this application dated October 10, 2006, and a second declaration dated October 14 2008. My scientific background and technical qualifications are set out in the first declaration.

2. The present invention is based, in significant part, on our finding that Guanine/Cytosine-enrichment (G/C enrichment) of mRNA sequences leads to stabilization of the mRNA inside cells, which prolongs the expression of the encoded protein in relation to a corresponding wild-type mRNA. Our studies indicate that this is a generic effect which is independent of the particular protein encoded by the mRNA. Stabilization of mRNA is one of the objects of the invention which is mentioned in our patent specification, e.g. at paragraph [0002]. Further experimental data is discussed in this declaration which supports our finding that G/C-enrichment increases mRNA stability. The following experiments were carried out under my supervision.

3. **Figure 1** attached hereto shows the level of luciferase expression *in vivo* after intramuscular injection of both wild type and G/C-enriched mRNA coding for *Photinus pyralis* luciferase ("Pp-Luciferase"). As seen in Figure 1, the luciferase

expression from the G/C-enriched mRNA was statistically significantly higher and, in particular , more extended in its duration as compared to expression from the wild type mRNA. The prolonged expression can only be explained by the G/C-enriched mRNA being more stable *in vivo* than the corresponding wild type mRNA.

4. A similar effect was observed in a homologous expression system (expression of a human protein in human cells), as shown in **Figure 2**. This figure reports results of the expression of interleukin-6 (IL-6) in HeLa cells at time periods 0-6h, 6-24h, 24-48h, and 48-72h post-transfection of both the wild type and the G/C-enriched mRNA coding for IL-6. As seen, there is no significant difference in the level of expression of IL-6 during the first 24 hours between the cells transfected with the G/C-enriched mRNA and the cells transfected with wild type mRNA, which means no significant difference in level of IL-6 expression at that time point. By the 3rd and 4th time points, however, there was a clear and dramatic difference in the expression levels. Whereas expression of IL-6 by the wild type mRNA decreased almost to the level of the control, expression by the G/C-enriched mRNA remained high. This experiment further supports our finding that stabilization of mRNA by G/C-enrichment causes prolonged (and, thereby, as a function of time higher) expression of the encoded protein by stabilizing the mRNA. The results of **Figure 1** and **Figure 2**, considered together, convincingly suggest that the stabilization of mRNA using G/C-enrichment is not dependent on the particular mRNA or the particular encoded protein product.

5. To my conviction, the "state of the art" at the time our patent application was filed provided no indication or expectation that G/C-enrichment of mRNA would lead to increased stability of the mRNA inside cells. The finding of enhanced mRNA stability was absolutely surprising to me and my coworkers, and I believe it would have been unexpected to any person working with mRNA at the time we made the invention. See, for example, Koide et al, "DNA Vaccines", Jpn. J. Pharmacol., 83: 167 (2000)(copy attached as **Exhibit D**), which states at page 168, 1st column, lines 4-7:

> "Although mRNA might be highly attractive owing to the lack of potential risk of integration into the genome, it does not seem very promising as a general method. The main reason seems to be the instability of the mRNA."

6. The invention as it is currently claimed in this patent application is directed specifically to a pharmaceutical composition which contains at least one G/C-enriched mRNA, wherein the G/C-enriched mRNA encodes a human tumor antigen. Since the pharmaceutical composition is for administration to a human, the human mRNA in the composition is for expression in human cells ("homologous" system). This can be contrasted with the situation where a (pathogenic) foreign gene is expressed (e.g. the expression of a bacterial gene in another non-bacterial cell system, e.g. in human cells) (a "heterologous" system).

7. My second declaration contains exemplary experimental data reporting our findings that G/C-enriched mRNA encoding human tumor antigens has improved anti-tumor activity *in vivo* as compared to the corresponding wild-type mRNA in a relevant mouse model. I again refer to that data, and like our findings with respect to mRNA stability, I am convinced the elevated anti-tumor activity of G/C-modified mRNA *in vivo* would not have been expected at the time of our invention.

8. In the most recent communication from the U.S. Patent and Trademark Office dated May 15, 2009, with which I am familiar, the Examiner has taken a position, *inter alia*: "A person of ordinary skill in the art would have readily recognized the benefit of optimizing the codons of any sequence for expression in a human cell regardless of the origin of the sequence (e.g. human, mouse, cow, bacteria, virus, ect.). One would have been motivated to make such a modification to improve the expression of the protein and reduce ribosomal pausing (e.g. Fomsgaard et al, page 1, lines 30-33)." As support in the prior art for the idea of "optimizing codons" the Examiner has cited to *Adema et al.* (US Pat. No. 6,500,919), *Fomsgaard* (WO 00/29561), and *Nagata et al*, J. Biochem. Biophys. Res. Comms. 261: 445 (1999).

9. I initially point out that increased mRNA stability and increased expression based upon codon optimization are distinct concepts. Our invention is based substantially on the unexpected finding of improved mRNA stability. I understand, however, that "codon optimization" may eventually but in no way necessarily result in an increased G/C content of DNA (and therefore of the mRNA transcribed therefrom), albeit for a different reason (increasing translation efficiency and not increasing mRNA stability).

10. It is correct that "optimization" of the codons of genes encoding pathogenic foreign genes to human codon usage may eventually result in an

increase in G/C content of the DNA (reflecting the "heterologous system").[1] Since the present claims recite G/C-enriched mRNA encoding human tumor antigens, one should ask whether, at the time of this invention, the skilled person would have appreciated any reason to increase the G/C content of a DNA encoding a human tumor antigen with an expectation of improving the expression of the encoded protein by that human DNA in a human cell ("the homologous system"). If there was no reason for the skilled person to G/C-enrich the DNA encoding human tumor antigens, it would follow that there was no reason to G/C-enrich the mRNA encoding human tumor antigens in the context of a pharmaceutical composition containing that mRNA.

11. The Adema et al. patent discloses the melanoma associated antigen gp100. The disclosure suggests vaccine use of the gp100 protein, its peptides and their nucleic acid sequences. Col. 7, lines 34-36. The disclosure suggests that, in addition to the sequence of SEQ ID NO: 1 (the naturally-occurring nucleic acid sequence encoding gp100 of SEQ ID NO: 2), variants of the sequence based on the degeneracy of the genetic code can be used. Col. 4, lines 50-60. The Adema et al. patent does not, however, specifically identify any G/C-enriched variants of the gp100 coding sequence, and does not suggest that any benefit(s) would be associated with using G/C-enriched variants of the gp100 coding sequence. The disclosure of Adema et al. merely reflects the common way to encompass the incredibly high number of conceivable sequence variants due to the degeneracy of the genetic code to generically expand the scope of protection as deliberately disclosed in almost any patent application filed in the field of DNA/protein sequences. The Fomsgaard disclosure concerns a DNA vaccine against Human Immunodefiency Virus (HIV). It is explained that one disadvantage of HIV envelope based DNA vaccines is their

---

[1] It is important to note, however, that modification to human codon usage does not necessarily increase G/C content. Like those of other organisms, human sequences exhibit a characteristic distribution of the frequency of the codons encoding for the 20 naturally occurring amino acids. Even though in human genes, some most preferred codons according to codon usage may be codons with a maximal possible G/C-content as it may occur in other cells and organisms, one must clearly distinguish between G/C-enrichment of an mRNA sequence and the codon usage optimization. As a specific example, the most preferred codon for the amino acid arginine in human cells is AGA although codons AGG, CGU, CGC, CGA and CGG, exhibiting a higher G/C-content, may be selected due to the degenerated genetic code (attached **Exhibit C**). A further specific example relates to serine. The most preferred codon for serine is AGC, but less preferred codons with the same G/C-content may also be selected, i.e. codons, which do not alter the G/C-content of the mRNA sequence.

intrinsically low expression which is specifically described in the scientific literature for HIV. Expression of HIV genes is regulated by HIV Rev protein. It is stated, with reference to an earlier paper to Haas (1996), that exchange of the HIV codon usage to the codon usage of highly expressed mammalian genes greatly improves the expression of DNA encoded HIV proteins in mammalian cell lines and renders the expression Rev-independent. Additionally, Fomsgaard mentions that "rare [HIV]codons" can cause pausing of the ribosome apparatus, which can interfere with translation. Page 1, lines 24-33.

12. There are several points to be made concerning the passage in the Fomsgaard document which was cited by the examiner, and the Fomsgaard reference as a whole. First, Fomsgaard's disclosure is concerned only with improving the expression of non-human, pathogenic HIV genes, which are known to be poorly-expressed in human cells ("heterologous system" in contrast to the present underlying concept of the present invention), so as to improve their expression in human cells and, thereby, increase their potential utility as anti-HIV vaccines. Second, the poorly-expressed HIV genes are expressed in a Rev-dependent manner and the desirability of obtaining their Rev-independent expression in human cells is recognized. Fomsgaard discloses codon optimization of an envelope gene from a primary HIV-1 clinical isolate. See WO 00/29561, p. 3, lines 1-8. The Haas et al. (1996) paper cited by Fomsgaard at p. 1, lines 27-30 reports codon optimization of the same HIV-1 envelope gene. Compare attached Haas paper **(Exhibit E)**, p. 315, col. 1, first paragraph, and p. 317, Fig. 2, with WO 00/29561, Figures 2-6. Third, there is no suggestion in Fomsgaard to optimize codons in non-HIV-1 genes, let alone human genes.

13. The examiner has cited Fomsgaard to support an assertion that one of skill in the art would have been motivated to codon optimize any gene with the expectation that increased expression of the encoded proteins would result. In fact, Fomsgaard does not suggest to codon-optimize "any gene." The skilled reader would recognize that Fomsgaard was continuing the work described in the Haas et al. paper, using exactly the same specific HIV gene. It is stated in the "Conclusion" section of the abstract in Haas et al.:

> "Codon-usage effects are a major impediment to the efficient expression of HIV-1 genes. Although mammalian genes do not

> show as profound a bias as do *Escherichia coli* genes, other ***proteins that are poorly expressed in mammalian cells can benefit from codon re-engineering***."

See **Exhibit E**, Haas et al., p. 322, column 1, "Conclusions" (emphasis added). As is evident, the concern of both Haas and Fomsgaard was restricted to the expression of HIV proteins, which are known to be poorly-expressed in human cells ("heterologous system" in striking contrast to the underlying concept of the present invention), and which are identified as the heterologous proteins which might benefit from "codon re-engineering."

14. The Examiner has also cited Fomsgaard to support an assertion that one of skill in the art would have been motivated to codon optimize any gene to minimize ribosomal pausing, which is suggested by Fomsgaard to result in premature termination of translation. See WO 00/29561, p. 1, lines 30-34. To my conviction, the purported effects of ribosomal pausing were not recognized by those of skill in the art at the time of our invention as being relevant to expression of <u>human</u> genes in <u>human</u> cells ("homologous system" according to the present invention). To my knowledge ribosomal pausing is a specific control mechanism primarily described for bacteria to react to amino acid deprivation which guarantees their survival and therefore is not a generic mechanism which is also present in mammalian cells as suggested by Fomsgaard. Such a control mechanism was, however, not known for the translation of mRNA in homologous systems, e.g. translation of human mRNA in human cells. In fact, the Haas et al. paper seems to dismiss the theory that an abundance of less-favored codons causes significant failure to complete a nascent polypeptide. See Haas et al., **Exhibit E**, p. 321, last paragraph. Accordingly, in my opinion, a concern about "rare [HIV] codons" to enhance expression in a "heterologous system" would not have been recognized by a skilled person as being relevant to a "homologous" system involving expression of genes encoding <u>human</u> tumor antigens in <u>human</u> cells.

15. It was logical to seek to improve the expression of poorly-expressed <u>pathogenic</u> ("heterologous") genes in human cells by modifying the pathogenic genes to utilize codons which are more prevalent in human cells as "heterologous" host system. Such an adaptation of codon usage, however, would not have been a concern when dealing with a homologous expression system. Specifically, the skilled

person would have seen no reason to introduce human codon usage into a human gene, such as a gene encoding a human tumor antigen.

16. Human tumor antigen genes are not (insofar as is known today) dependent on HIV Rev protein for their expression. Fomsgaard's disclosure which relates to rendering expression of HIV genes Rev-independent has no applicability to a homologous expression system (i.e. a pharmaceutical composition containing mRNA encoding human tumor antigen genes for administration to a human (as disclosed by the present invention).

17. Nagata et al. experimented with small subsequences of DNA sequences derived from bacterial and protozoal pathogens and adapted the codon usage of those subsequences to the codon usage of mammalian cells and tested the expression therefrom and induction of cytotoxic T lymphoctes (CTL) once again reflecting the "heterologous system". A starting point for the experiments was the reccognition that "a variety of bacterial and protozoa utilize a biased codon usage different from that in mammalian genes." Page 445, 1st column. The results suggested that the polypeptide expression level and CTL induction were higher upon substitution of "optimized" codons. Page 450, 2d column, last 7 lines. In the specific context being described – replacing the pathogen's ("heterologous") codons with human-preferred codons – the "optimizing" of the codons is equivalent to "humanizing" the codons for "heterologous" expression in a human system. The assumption is that human codon usage would be "optimal" for expression of pathogenic foreign proteins in human cells. There is not even a suggestion in Nagata that optimizing/humanizing would be a relevant concern when seeking to express human genes in human cells.

18. Nagata co-authored a review article published in 2000 (one year later than the cited Nagata *et al.* paper, which is listed as reference 19 in the review). Koide et al (**Exhibit D**). In the discussion of codon "optimization" on page 171, 2d column, the following passage appears:

> "Although many bacteria have been targets of DNA vaccines, significant progress has been made with *Mycobacterium tuberculosis*. One of the reasons seems to be the bias in codon usage in *M. tuberculosis* genes. Surprisingly, the bias, unlike any other bacteria examined, is comparable to that of *Mus musculus* and *Homo Sapiens* (17), suggesting that the codon-optimization

described above could be unnecessary for the construction of DNA vaccines against M. tuberculosis."

This statement reflects that, according to Nagata and his co-authors - who were actually working with "codon optimization" experiments at the relevant time prior to our invention - "codon optimization" was not considered to be relevant or necessary in the case of genes already having a mammalian codon bias as e.g. encountered for M. tuberculosis.

19. Consistent with what was stated by Haas et al (quoted above) the statement from Nagata's review article is indicative that the need to "optimize" codon usage was not recognized as a "problem" in the prior art except as it related to preparation of DNA vaccines to immunize humans against pathogenic organisms (the focus of both Fomsgaard and Nagata) or "heterologous" genes poorly expressed in mammalian cells. See also, Wu *et al.*, Mol. Ther., 2: 288 (2000), (abstract attached as **Exhibit F**; suggesting need to codon-optimize or otherwise engineer for improved expression in relevant only to "heterologous" (HIV) genes which are "difficult to translate by mammalian cells"). For example, there would be no reason why a human gene would need to be optimized/humanized by modification to introduce human codon usage, since it reflects by its very nature human codon usage. Indeed, assuming a need to optimize human genes for expression in human cells would have been counterintuitive, since the skilled person would have believed that millions of years of evolution would have fine-tuned human gene expression in human cells for more efficiently than could be achieved by designing man-made, synthetic nucleic acid sequences. Moreover, the belief that codon usage was understood to influence the expression level in human genes at the time of our invention is unfounded. I am not aware that any such correlation had been established or was appreciated by persons in the art. See, for example, Duret et al, PNAS USA, 96: 4482 (1999) (**Exhibit G**) at page 4487, col. 1, last full paragraph ("...we did not observe any correlation between codon usage and expression level in human genes (33) (unpublished data). As noticed by others (6, 9), this absence of selection may be explained by population genetics ...").

20. I further note that the cited references to Fomsgaard and Nagata disclose the codon-optimization of DNA, not mRNA as disclosed by the present

invention. While changes to DNA would be reflected at the RNA level if and when the DNA is properly transcribed in a cell, the claims here concern a pharmaceutical composition containing modified mRNA. Neither of the references to Fomsgaard or Nagata suggests to modify an mRNA and to provide the modified mRNA in the isolated environment of a pharmaceutical composition.

21. For the above reasons, the skilled artisan would not have expected the increased stability of a G/C-enriched mRNA in relation to a wild type mRNA and even I – as an inventor_- was perfectly surprised, when I analyzed the original experimental findings which then led to the present invention. To the extent that "codon optimization" would lead to G/C-enriched DNA and therefore RNA when transcribed, there was no recognized need (or logical reason) in the art as to why modify codons of human genes. Since there was no reason to modify human genes for human expression ("homologous system"), nor any expectation of increased expression of human genes in a human system, the improved *in vivo* anti-tumor activity which is achieved using G/C-enriched mRNA encoding human tumor antigens in a mouse model could not have been reasonably predicted.

22. I want to finally further emphasize that clinical phase trials are currently ongoing and initiated by the assignee of the present application which are directed to tumor vaccination of prostate cancer patients and non-small cell lung cancer patients, who are treated with G/C enriched mRNA encoding human tumor antigens (according to the present invention). Clinical phase I trials could show safety of the mRNA vaccines and clinical results will be expected next year. So far preclinical data are extremely encouraging.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of this application of any patent issued thereon.

Date: 10 12 04

Dr. Ingmar Hoerr

Encl.: Exhibit A: Figure 1
Exhibit B: Figure 2
Exhibit C: Codon Usage Databased, and chart from Wikipedia
Exhibit D: Koide et al, "DNA Vaccines", Jpn. J. Pharmacol., 83: 167 (2000)
Exhibit E: Haas, et al, Current Biology, 6: 315 (1996)
Exhibit F: Wu *et al.*, Mol. Ther., 2: 288 (2000), (abstract only)
Exhibit G: Duret et al, PNAS USA, 96: 4482 (1999)

Exhibit A:

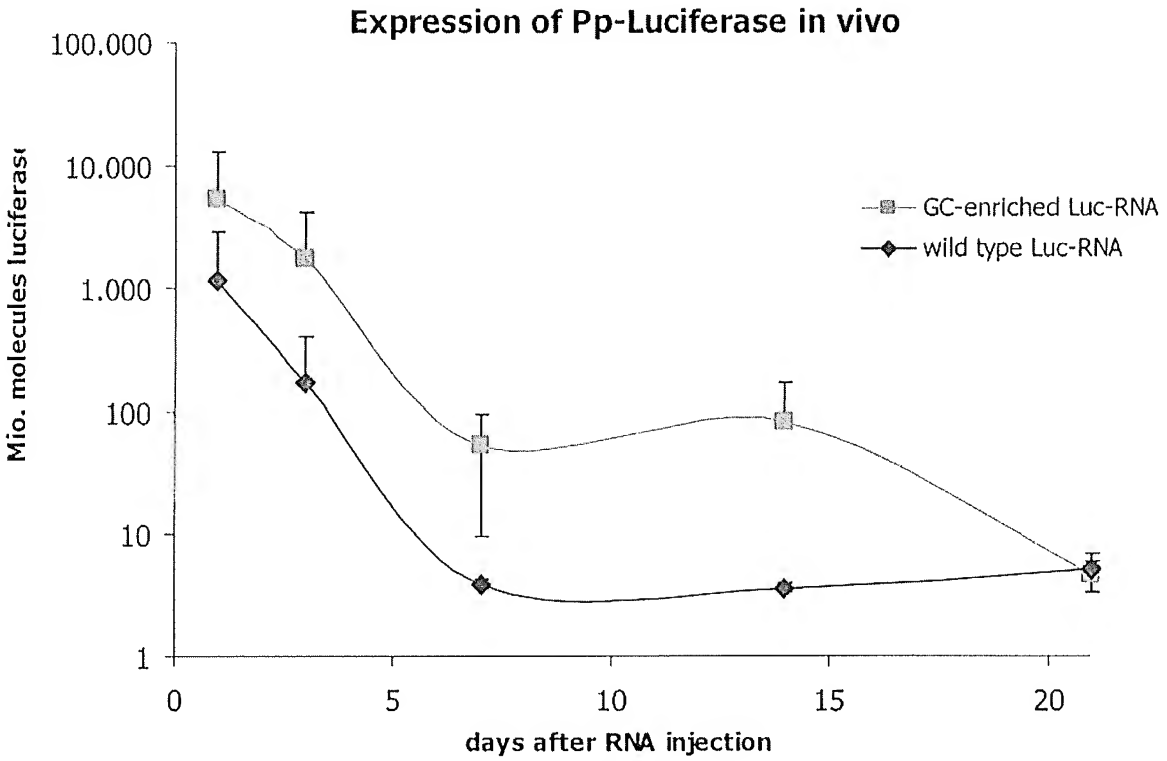


Fig.1: Expression of luciferase after intramuscular injection of wild type or Guanine/Cytosine-enriched mRNA coding for Photinus pyralis Luciferase (Pp Luc) in mice. Luciferase expression was measured after the indicated time.

Exhibit B

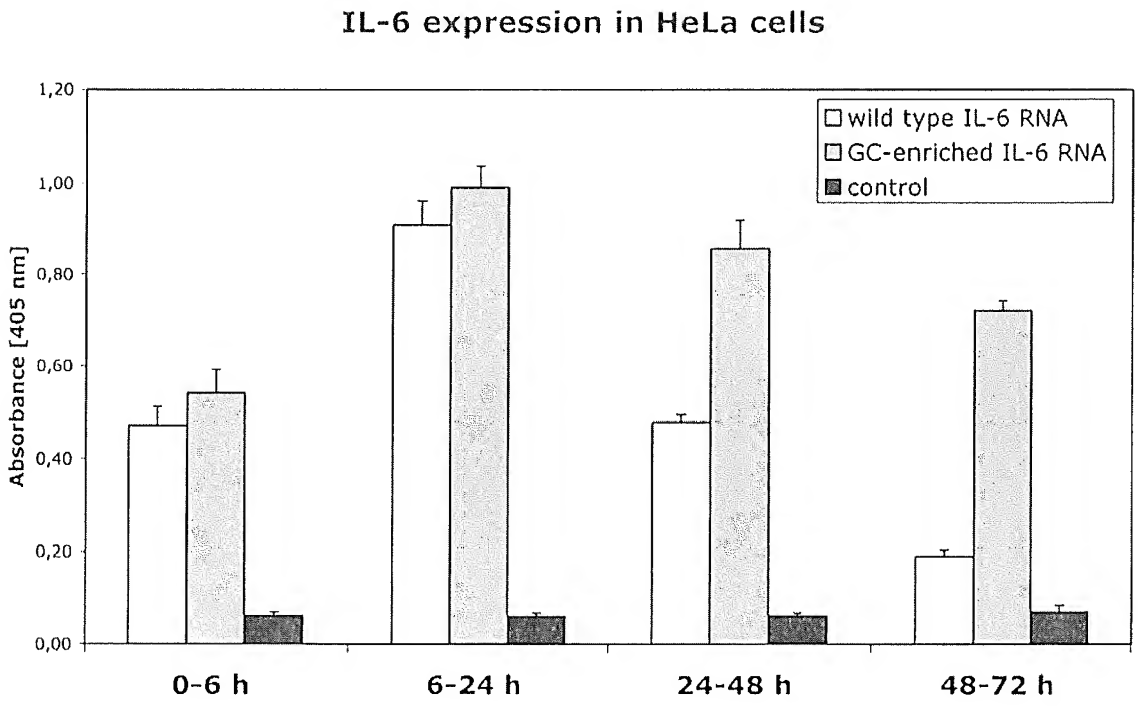


Fig. 2 IL-6 expression in HeLa cells after transfection with wild type or Guanine/Cytosine-enriched mRNA coding for interleukin 6 (IL-6). Supernatants of the HeLa cells were measured for IL-6 expression for the indicated time periods after transfection.

Exhibit C

Source : Codon Usage Databased (www.kazusa.or.jp/condon/)

***Homo sapiens* [gbpri]: 93487 CDS's (40662582 codons)**

fields: [triplet] [frequency: **per thousand**] ([number])

UUU 17.6(714298)	UCU 15.2(618711)	UAU 12.2(495699)	UGU 10.6(430311)
UUC 20.3(824692)	UCC 17.7(718892)	UAC 15.3(622407)	UGC 12.6(513028)
UUA 7.7(311881)	UCA 12.2(496448)	UAA 1.0(40285)	UGA 1.6(63237)
UUG 12.9(525688)	UCG 4.4(179419)	UAG 0.8(32109)	UGG 13.2(535595)
CUU 13.2(536515)	CCU 17.5(713233)	CAU 10.9(441711)	CGU 4.5(184609)
CUC 19.6(796638)	CCC 19.8(804620)	CAC 15.1(613713)	CGC 10.4(423516)
CUA 7.2(290751)	CCA 16.9(688038)	CAA 12.3(501911)	CGA 6.2(250760)
CUG 39.6(1611801)	CCG 6.9(281570)	CAG 34.2(1391973)	CGG 11.4(464485)
AUU 16.0(650473)	ACU 13.1(533609)	AAU 17.0(689701)	AGU 12.1(493429)
AUC 20.8(846466)	ACC 18.9(768147)	AAC 19.1(776603)	AGC 19.5(791383)
AUA 7.5(304565)	ACA 15.1(614523)	AAA 24.4(993621)	AGA 12.2(494682)
AUG 22.0(896005)	ACG 6.1(246105)	AAG 31.9(1295568)	AGG 12.0(486463)
GUU 11.0(448607)	GCU 18.4(750096)	GAU 21.8(885429)	GGU 10.8(437126)
GUC 14.5(588138)	GCC 27.7(1127679)	GAC 25.1(1020595)	GGC 22.2(903565)
GUA 7.1(287712)	GCA 15.8(643471)	GAA 29.0(1177632)	GGA 16.5(669873)
GUG 28.1(1143534)	GCG 7.4(299495)	GAG 39.6(1609975)	GGG 16.5(669768)

Coding GC 52.27% 1st letter GC 55.72% 2nd letter GC 42.54% 3rd letter GC 58.55%

Source: Wikipedia (http://de.wikipedia.org/wiki/Codon_Usage#cite_note-0)

Aminosäure	Codon	*E. coli* K12	*S. cerevisiae*	*H. sapiens*	Aminosäure	Codon	*E. coli* K12	*S. cerevisiae*	*H. sapiens*
Valin (V)	GUU	18,8	22,1	11,0	Alanin (A)	GCU	10,7	21,2	18,4
	GUC	11,7	11,8	14,5		GCC	31,8	12,8	27,7
	GUA	11,5	11,8	7,1		GCA	21,1	16,2	15,8
	GUG	26,4	10,8	28,1		GCG	38,5	6,2	7,4
Leucin (L)	CUU	11,9	12,3	13,2	Prolin (P)	CCU	8,4	13,5	17,5
	CUC	10,5	5,4	19,6		CCC	6,4	6,8	19,8
	CUA	5,3	13,4	7,2		CCA	6,6	18,3	16,9
	CUG	48,9	10,5	39,6		CCG	26,7	5,3	6,9
Leucin (L)	UUA	15,2	26,2	7,7	Serin (S)	UCU	5,7	23,5	15,2
	UUG	11,9	27,2	12,9		UCC	5,5	14,2	17,7
Phenylalanin (F)	UUU	19,7	26,1	17,6		UCA	7,8	18,7	12,2
	UUC	15,0	18,4	20,3		UCG	8,0	8,6	4,4
Isoleucin (I)	AUU	30,5	30,1	16,0	Threonin (T)	ACU	8,0	20,3	13,1
	AUC	18,2	17,2	20,8		ACC	22,8	12,7	18,9
	AUA	3,7	17,8	7,5		ACA	6,4	17,8	15,1
Methionin (M)	AUG	24,8	20,9	22,0		ACG	11,5	8,0	6,1
Aminosäure	Codon	*E. coli* K12	*S. cerevisiae*	*H. sapiens*	Aminosäure	Codon	*E. coli* K12	*S. cerevisiae*	*H. sapiens*
Asparaginsäure (D)	GAU	37,9	37,6	21,8	Glycin (G)	GGU	21,3	23,9	10,8
	GAC	20,5	20,2	25,1		GGC	33,4	9,8	22,2
Glutaminsäure (E)	GAA	43,7	45,6	29,0		GGA	9,2	10,9	16,5
	GAG	18,4	19,2	39,6		GGG	8,6	6,0	16,5
Tyrosin (Y)	UAU	16,8	18,8	12,2	Cystein (C)	UGU	5,9	8,1	10,6
	UAC	14,6	14,8	15,3		UGC	8,0	4,8	12,6
Stopp	UAA	1,8	1,1	1,0	Stopp	UGA	1,0	0,7	1,6
Stopp	UAG	0,1	0,5	0,8	Tryptophan (W)	UGG	10,7	10,4	13,2
Asparagin (N)	AAU	21,9	35,7	17,0	Serin (S)	AGU	7,2	14,2	12,1
	AAC	24,4	24,8	19,1		AGC	16,6	9,8	19,5
Lysin (K)	AAA	33,2	41,9	24,4	Arginin (R)	AGA	1,4	21,3	12,2
	AAG	12,1	30,8	31,9		AGG	1,6	9,2	12,0
Histidin (H)	CAU	15,8	13,6	10,9	Arginin (R)	CGU	21,1	6,4	4,5
	CAC	13,1	7,8	15,1		CGC	26,0	2,6	10,4
Glutamin (Q)	CAA	12,1	27,3	12,3		CGA	4,3	3,0	6,2
	CAG	27,7	12,1	34,2		CGG	4,1	1,7	11,4

Jpn. J. Pharmacol. 83, 167 – 174 (2000)

Exhibit D

REVIEW —*Current Perspective*—

DNA Vaccines

Yukio Koide, Toshi Nagata, Atsushi Yoshida and Masato Uchijima

Department of Microbiology and Immunology, Hamamatsu University School of Medicine, Hamamatsu 431-3192, Japan

Received March 21, 2000

ABSTRACT—DNA vaccination or genetic immunization is a rapidly developing technology that offers new approaches for the prevention and therapy of disease. Regarding the inoculation method of DNA vaccine, we recommend the gene gun delivery system, which is a highly reliable method compared to intramuscular inoculation. DNA vaccines could have potential advantages over other types of vaccines in that these vaccines can induce strong cellular immune responses, cytotoxic T lymphocytes and type 1 helper T cells, without resorting to live organisms or complicated protein formulation. The cellular immune responses are especially required for the protection against infections with intracellular pathogens such as viruses and *Mycobacterium tuberculosis* and protection against cancers, suggesting that they seem to be suitable targets of DNA vaccines. We describe here that their application to bacterial infections requires optimization of codon usage in the DNA vaccines to the host animal to improve translational efficiencies of the bacteria genes. DNA vaccines for a variety of pathogens and cancers have now entered phase I/II human clinical trials.

Keywords: DNA vaccine, Plasmid, Cytotoxic T lymphocyte, CpG motif

Introduction

Since Jenner introduced vaccinia virus inoculation against smallpox two centuries ago, vaccination against infectious diseases has had a long and successful history. Vaccines have eradicated smallpox and pushed polio virus to the brink of extinction. Thus, active immuno-prophylaxis through vaccination has become the most effective and cost-effective public health measure available. Today, diverse vaccines are available, including 1) live attenuated vaccines (e.g., *Mycobacterium bovis* BCG, Sabin vaccine for poliomyelitis), 2) killed or inactivated vaccines (e.g., cholera vaccine, Rocky Mountain spotted fever vaccine for rickettsial infection), 3) component vaccines (e.g., influenza HA vaccine, acellular *Bordetella pertussis* vaccine), 4) recombinant vaccines (e.g., major surface protein of hepatitis B virus), 5) toxoid vaccines (formalin-detoxified diphtherial toxin, tetanal toxoid). Many of these vaccines induce neutralizing antibodies. As efficient protection against intracellular pathogens critically depends on cellular immune responses, inactivated and soluble protein vaccines may be insufficient and live attenuated vaccines are considered essential. However, the live attenuated vaccines will inevitably suffer from the problems associated with adverse reactions where the host immune response participates in the attenuated phenotype.

Vaccination with plasmid DNA has potential advantages compared to these traditional protein vaccinations due to the strong cellular immune responses induced in addition to humoral immune response (1, 2). Therefore, DNA vaccination can serve as an alternative to immunization with the live vaccines, which are essential for the protection against infections with intracellular bacteria. Two major arms of cellular immunity can come into play in the protection. Type 1 helper T (Th1) cells play a pivotal role in the protection against infections with intracellular bacteria such as *Mycobacteria spp.* or *Salmonella spp.* which persist in the endosome, while cytotoxic T lymphocytes (CTL) eradicates pathogens persisting in the cytosole such as viruses, *Rickettsia spp.*, or *L. monocytogenes*. Both cellular immune responses have been shown to be effectively induced with DNA vaccines.

Plasmid DNA-mediated immunization

DNA vaccines may be in the form of either DNA or mRNA. However, almost all animal studies to date have used *Escherichia coli*-derived plasmid DNA, which is composed of an antigen-encoding gene whose expression is regulated by a strong mammalian promoter such as cytomegalovirus immediate-early promoter/enhancer (Fig. 1). The plasmid also possesses a polyadenylation termination

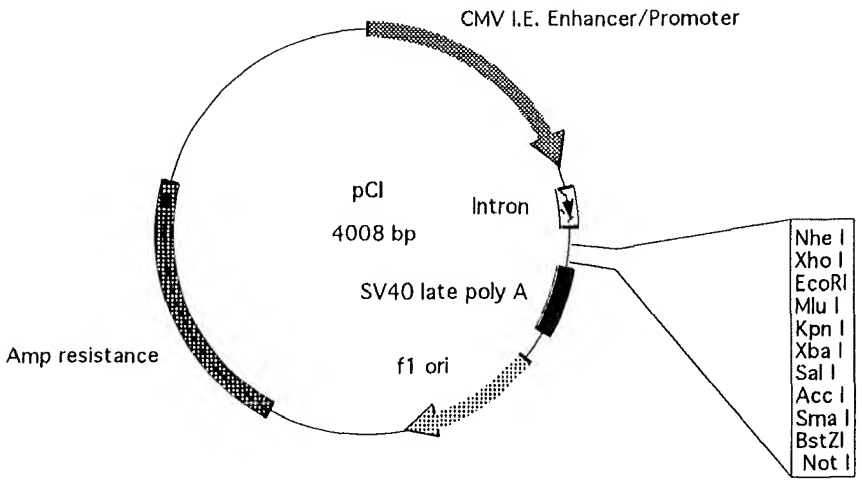


Fig. 1. Plasmid DNA derived from *E. coli* forms the basis of DNA vaccines and optimally contains the following components: a strong eukaryotic promoter/enhancer such as cytomegalovirus immediate-early promoter/enhancer, a multiple cloning site, a polyadenylation sequence (e.g., SV40 late poly A), a prokaryotic origin of replication (*ori*), and a selectable marker (e.g., ampicillin resistance gene [*amp*]).

sequence, a prokaryotic origin of replication (*ori*) in *E. coli* and a selective marker such as ampicillin resistance gene (*amp*) to facilitate selection of cells harboring the plasmid. Although mRNA might be highly attractive owing to the lack of potential risk of integration into the genome, it does not seem very promising as a general method. The main reason seems to be the instability of mRNA. However, Ying et al. (3) described cancer therapy using a self-replicating RNA vaccine. Employing a gene encoding an RNA replicase polyprotein derived from the Semliki forest virus, they demonstrated that a single intramuscular injection of the self-replicating RNA immunogen is capable of inducing a specific antibody and $CD8^+$ T cell responses at doses as low as 0.1 μg.

Immunization with DNA can be accomplished by two fundamentally different ways. One approach is needle injection into different tissue sites, with the most common route being intramuscular injection into the hid leg quadriceps or tibialis anterior. Alternatively, plasmid DNA can be administrated by a gene gun to propel the DNA-coated gold particles into the epidermis. In the former case, the mechanism by which plasmid DNA enters the muscle cells is not known. However, T-tubules found exclusively in striated muscle have been implicated in uptake of plasmid DNA since they are involved in the invagination of the plasmid membrane. A much greater improvement in efficacy of gene transfer can be achieved by injection of plasmid DNA into regenerating skeletal muscle. Muscle degeneration/regeneration can be induced by use of cardiotoxin (4) or local anesthetics such as bupivacaine (5).

It is of particular interest that gene gun DNA immunization requires 100- to 1000-fold less DNA than muscle DNA inoculation to generate an equivalent antibody response (6). For both needle muscular injection of plasmid DNA and gene gun epidermal immunization, bone marrow-derived antigen presenting cells (APC) were demonstrated to be responsible for the presentation of the expressed antigen to precursor CTL restricted to the donor MHC haplotype in F1 bone marrow chimeric mice (7). In the case of gene gun DNA immunization, Langerhans cells within the epidermis or even dermal macrophages have been suggested to act as APC (8). Of particular interest is that excision of an injected muscle bundle within 10 min of DNA inoculation does not affect the magnitude or longevity of antigen-specific antibody responses. By contrast, biopsy of the skin target site up to 24 h after gene gun bombardment completely abrogated the antibody response in the majority of mice (9), suggesting that transfected cells in gene gun-bombarded skin, but not needle-injected muscle, play a central role in DNA-initiated antibody and CTL responses. These data imply that muscle cells have nothing to do with intramuscular DNA immunization and that APC infiltrating into muscles are responsible for transcription and translation of the DNA vaccines as well as processing and presentation of the antigens.

It has been suggested that muscle DNA immunization raises a predominantly Th1 response, while gene gun DNA immunization produces a Th2 response (10). The shift to the Th1 response in muscle DNA immunization is considered to be mediated by CpG motifs present in plasmid DNA vaccines. It is generally accepted that a CpG motif consisting of an unmethylated CpG dinucleotide flanked by two 5′ purines (optimally a GpA) and two 3′ pyrimidines (optimally a TpC or TpT) stimulated the innate immune

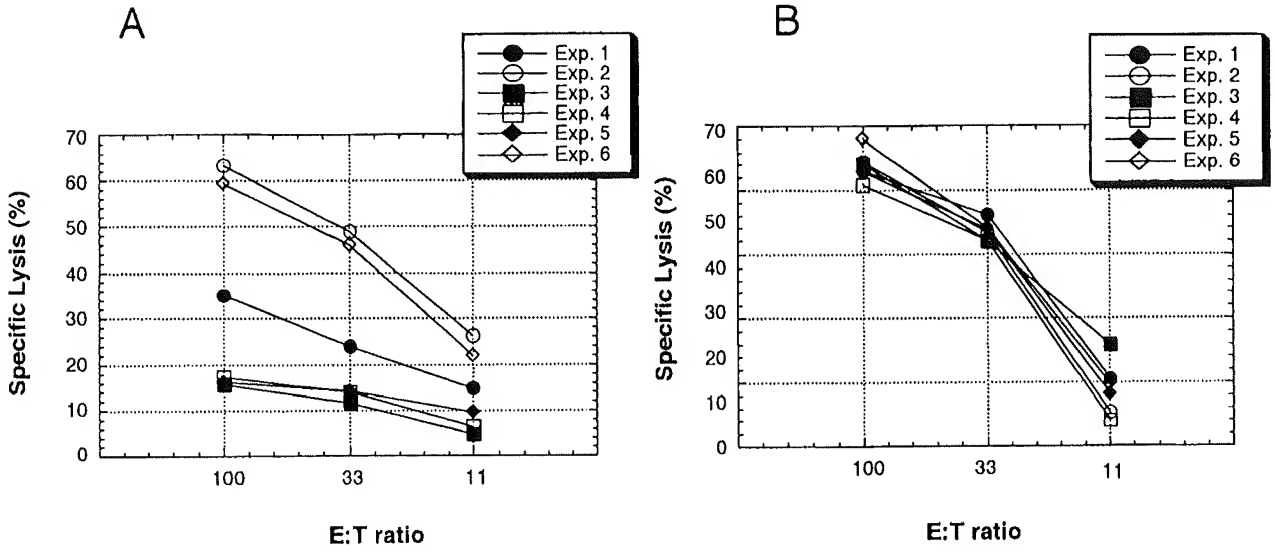


Fig. 2. Comparison of CTL activity specific for LLO 91–99 by intramuscular and gene gun-mediated immunization. BALB/c mice were immunized either with 50 μg of pCI-OVA intramusculary (A) or with 2 μg of pCI-OVA by the gene gun system (B) three times. Splenocytes from immunized mice were harvested 2 weeks after the last immunization and stimulated in vitro with LLO 91–99-pulsed splenocytes for 5 days. Percentage of specific lysis was determined using J774 cells (H-2^d) pulsed with LLO 91–99 peptide as target cells. Results are expressed as the mean for triplicate determinations.

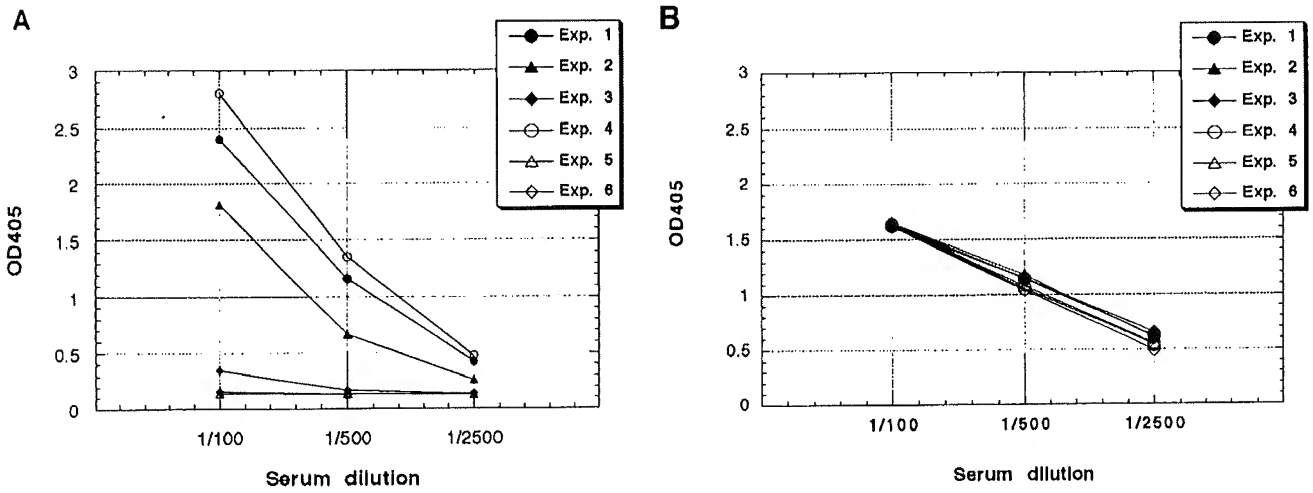


Fig. 3. Comparison of anti-OVA IgG antibody induction by intramuscular and gene gun-mediated immunization. BALB/c mice were immunized with either 50 μg of pCI-OVA intramusculary (A) or 2 μg of pCI-OVA by the gene gun system (B), three times. Three weeks after the last immunization, sera were obtained and analyzed for the presence of OVA-specific antibodies by ELISA. Serum dilutions were 1/100, 1/500 and 1/2500. OD405, optical density at 405 nm.

system to produce a series of immunomodulatory cytokines (11). The cytokines include interleukin (IL)-12 and interferon (IFN)-γ (12), which promote the development of Th1 cells. Vertebrate DNA lacks these immune effects. However, the mechanism by which gene gun DNA immunization induces the Th2 response remains elusive. The signaling pathway elicited in monocytes and B cells by the CpG motif was investigated (13). CpG DNA may be taken up by leukocytes via adsorptive endocytosis into an acidified chloroquine-sensitive intracellular compartment. The endosomal acidification of the CpG DNA is required for the CpG DNA-mediated leukocyte activation and is coupled to the rapid generation of reactive oxygen species (ROS). These ROS may be required for IκBα and IκBβ phosphorylation and degradation and subsequent NFκB activation leading to cytokine gene expressions.

Utilizing a plasmid DNA encoding a single CTL epitope and that encoding ovalbumin (OVA), we have compared the reproducibility in the induction of CTL and antibody by gene gun and intramuscular immunization. As shown in Fig. 2A, the results of CTL assay by the intramuscular DNA injection varied significantly. Only two experiments among a total of six showed high specific CTL activity, whereas CTL activities obtained with gene gun immunization are very reproducible (Fig. 2B). The same is true in antibody production. As shown in Fig. 3A, intramuscular injection of DNA vaccine encoding OVA induced significant level of serum anti-OVA IgG in only three of six mice. In contrast, all mice immunized by gene gun produced significant and constant levels of anti-OVA IgG (Fig. 3B). Thus, we believe that the gene gun delivery system is a highly reliable method compared to intramuscular inoculation (14).

Animal models of DNA vaccine

A substantial number of experimental models of DNA vaccination have been reported (Table 1). Since immunization of DNA vaccine resembles a virus infection, most of the pathogens studied have been viruses. One of the first uses of DNA vaccine was to induce CTL against infection with influenza A virus. Ulmer et al. (15) reported that intramuscular immunization of DNA vaccine encoding influenza A nucleoprotein (NP) successfully induced CTL and that the CTL could protect mice against the viral infections even when the influenza virus is carrying a different hemagglutinin but the same NP.

Hepatitis B virus (HBV) infects only human and chimpanzees, and consequently, a challenge with the virus cannot be performed in other species. Therefore, several experiments with HBV DNA vaccine have been performed employing chimpanzees. However, since antibodies alone are able to provide protective immunity to the infection, murine DNA vaccination to HBV is a relevant model of a potent human DNA vaccine. Intramuscular immunization with HBV surface antigen (HBsAg) was demonstrated to induce CTL and Th1 responses as well as the humoral response (4). Responses follow the same pattern when mice are injected with DNA encoding other viruses such as human immunodeficiency virus (HIV) (16).

DNA vaccines encoding bacteria have also been used with success. Since DNA vaccines have advantages over conventional protein vaccines in that they can induce strong cellular responses, the vaccine seems to be most efficient at the protection against infections with intracellular bacteria among bacterial infections.

We have been working on DNA immunization for *L. monocytogenes*, a Gram-positive facultative intracellular bacterium. The bacterium is known to induce major histocompatibility complex (MHC) class I-restricted $CD8^+$ T cell responses in addition to MHC class II-restricted

Table 1. DNA vaccines successfully used in animal models

Vaccines against	Proteins encoded by DNA vaccines	Results		
		antibodies	CTL	protection
HIV	Env, Gag, Rev	+	+	ND
Influenza virus	NP, HA, M1	+	+	+
Hepatitis B virus	HBs, Core antigens	+	+	+
Hepatitis C virus	Core/Nucleocapsid	+	+	ND
Herpes simplex virus	GB, gD, ICP27	+	+	+
Papillomavirus	L1	+	ND	+
HTLV-1	Env	+	ND	ND
Bovine herpes virus	gp	+	ND	ND
Rabies virus	Gp, NP	+	+	+
Lymphocytic choriomeningitis virus	NP	+	+	+
Plasmodium sp	CSP	+	+	+
Leishmania major	gp63	+	ND	+
Mycobacterium tuberculosis	HSP65, Ag85, MPT70	+	+	+
Mycoplasma pulmonis	A7-1, A8-1	+	+	+
Salmonella typhi	OmpC porin	+	ND	ND
Listeria monocytogenes	LLO, p60	ND	+	+
Bacillus thuringiensis	Endotoxin	+	ND	ND
Chlamydia trachomatis	MOMP	+	+	+
Tetanus toxin	Fragment C	+	ND	ND
Borrelia burgdorferi	OspA	+	+	+
Melanoma	MAGE-1, MAGE-3	ND	+	+
B-cell lymphoma	Single chain Fv (idiotope)	+	ND	+
Renca, MethA, SA-1, L5178V, p815	IL-12	ND	+	+

ND, not determined.

$CD4^+$ T cell responses since *L. monocytogenes* is able to escape from the phagocytic vesicle into the cytosol of the infected cells, thereby introducing bacterial proteins into the MHC class I antigen processing pathway. Therefore, murine infection with *L. monocytogenes* is a widely used model for studying MHC class I-restricted $CD8^+$ T cell responses against intracellular bacteria. There have emerged four different *L. monocytogenes* epitopes presented by MHC class I H-$2K^d$ molecules to CTL: listeriolysin O (LLO) 91–99, p60 217–225, p60 449–457, and mpl 84–92. Two of these four epitopes, LLO 91–99 and p60 217–225, are demonstrated to induce dominant responses. We, therefore, constructed a plasmid DNA vaccine expressing the wild type DNA sequence of LLO 91–99 (p91wt). However, inoculation of p91wt into BALB/c mice failed to induce CTL against the nonapeptide-pulsed J774 target cells (H-2^d). We hypothesized that one of the reasons for the failure is the poor translation efficiency of the p91wt mRNA in the murine cell environment since the bias in codon usage observed in many species affects the translation efficiency of selective codons in a given gene. We, therefore, assessed the difference of codon bias in genes between *L. monocytogenes* and *Mus musculus*. For the purpose, we constructed a reference table of relative synonymous codon usage (RSCU) values from very highly expressed genes of the organism in question. An RSCU value for a codon is simply the observed frequency of that codon divided by the frequency expected under the assumption of equal usage of the synonymous codons for an amino acid. As expected, the bias in codon usage in *L. monocytogenes* genes appeared to be totally different from that in *Mus musculus* (17). Therefore, we constructed a plasmid expressing LLO 91–99, p91mam, in which native codons were substituted with codons frequently found in highly expressed murine genes. As an experimental approach to compare translation efficiencies between p91wt and p91mam, we employed "read-through analysis" in which the luciferase cDNA was fused to downstream of the wild type or adapted LLO 91–99 sequence, resulting in p91wt-Luc and p91mam-Luc, respectively, to express LLO 91–99/Luciferase fusion proteins; and a conventional luciferase assay was performed using BALB/3T3 murine fibroblast cells. In this experiment, luciferase activities are critically dependent on the translation efficiencies of the upstream sequences, LLO 91–99wt and LLO 91–99mam. The relative luciferase activities of p91mam-Luc were remarkably higher than those of p91wt-Luc in BALB/3T3 cells, suggesting that p91mam expressed a much higher level of LLO 91–99 in the immunized murine cells compared to p91wt (18). The codon-optimized p91mam successfully induced CTL against the nonapeptide-pulsed J774 cells (Fig. 4) (18, 19). In order to determine the biological effect of 91mam and p91wt in combating bacterial challenge, BALB/c mice immunized with p91wt or p91mam were challenged with a sublethal dose of *L. monocytogenes*, and the colony forming unit (CFU) from the spleen and liver was counted. Mice immunized with p91mam have 1 to 2 logs lower CFU in the spleen and liver than control mice. On the contrary, mice immunized with p91wt showed the number of CFU comparable to that in control mice. The data indicate that p91mam vaccine is effective in controlling the organism replication (18).

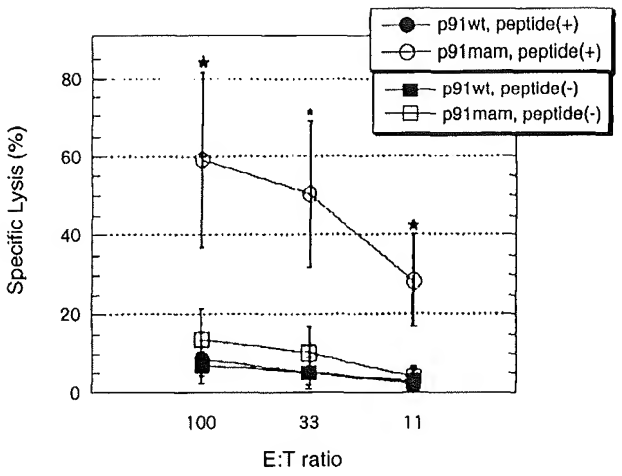


Fig. 4. Codon usage effects of DNA vaccine on CTL induction. BALB/c mice were immunized with p91wt (●, ■) or p91mam (○, □), three times biweekly. Spleen cells from immunized mice were harvested 2 weeks after the last immunization and stimulated in vitro with LLO 91-99-pulsed splenocytes for 5 days. Percentage of specific lysis was determined using J774 cells (H-2^d) pulsed with LLO 91–99 peptide (●, ○) or control medium (■, □) as target cells. Results are expressed as the mean for triplicate determinations. *Statistically significant differences in specific lysis between p91mam against J774 pulsed with LLO 91–99 and the others ($P<0.005$).

Although many bacteria have been targets of DNA vaccines, significant progress has been made with *Mycobacterium tuberculosis*. One of the reasons seems to be the bias in codon usage in *M. tuberculosis* genes. Surprisingly, the bias, unlike any other bacteria examined, is comparable to that in *Mus musculus* and *Homo sapiens* (17), suggesting that the codon-optimization described above could be unnecessary for the construction of DNA vaccines against *M. tuberculosis*.

Several protective antigens of *M. tuberculosis* have been revealed: 65-kDa heat shock protein (hsp65), hsp70, Ag85, 38-kDa Ag and 6-kDa early secretary antigenic target (ESAT-6). Therefore, plasmid DNA encoding these antigens could be employed as DNA vaccines for tuberculosis. In 1996, two separate papers appeared to show the efficacy of DNA vaccines in the protection against tuberculosis. Tascon et al. (20) showed that immunization with plasmid DNA encoding hsp65 induced strong humoral and cellular immune responses. It is noteworthy that this DNA vaccine

could protect mice against i.p. challenge with *M. tuberculosis* and that the protection was equivalent to that obtained using *M. bovis* BCG. Huygen et al. (21) reported that immunization of mice with plasmid DNA encoding Ag85 complex composed of Ag85A, Ag85B and Ag85C induced high titer antibodies of both IgG1 and IgG2 and IFN-γ production. In further experiments, immunization of the DNA vaccine conferred protection against challenge with *M. tuberculosis*. Lozes et al. (22) also studied the plasmid DNA vaccines encoding each subcomponent of Ag85 complex and found that the DNA vaccines encoding Ag85A or Ag85B were able to induce Th1 type immune responses and CTL against *M. tuberculosis*, resulting in protective immune responses. In contrast, the DNA vaccine encoding Ag85C failed to do so. Likewise, Montgomery et al. (23) reported the success of a DNA vaccine encoding Ag85A in the tuberculosis challenge model. A substantial number of published reports have indicated the efficacy of DNA vaccines encoding hsp65 in protection against tuberculosis. In addition, a DNA vaccine encoding 38-kDa Ag was demonstrated to induce a mycobacterium-specific Th1 response and CTL response. Again, protection obtained by the vaccine was equivalent to that seen with BCG vaccination.

DNA vaccine to *M. tuberculosis* was also demonnstrated to have a pronounced therapeutic action. Lowrie et al. (24) demonstrated that DNA vaccines encoding Hsp65 and MPT70 showed strong therapeutic effect against the infection. Furthermore, DNA vaccine therapy has been shown to be equally effective against an established infection with a drug-resistant strain of *M. tuberculosis*.

Increasing evidence suggests that cell-mediated immunity, particularly T cell-mediated immunity, is important for the control of tumor cells. Therefore, DNA vaccine seems to be a promising strategy to enhance cell mediated immunity against tumors. MAGE-1 and MAGE-3 are two clinically relevant antigens expressed in many human melanomas and other tumors, but not in normal tissues, except testis. Intramuscular expression of MAGE-1 and MAGE-3 by plasmid DNA injection and subcutaneous immunization with syngeneic mouse embryonic fibroblasts transduced with recombinant retroviruses to express these antigens induced specific immunity against tumors expressing MAGE-1 and MAGE-3 (25). IL-12 plays an important role in the induction of protective immunity against cancer. Intradermal injection of IL-12 cDNA resulted in local expression of IL-12 mRNA, which correlated with a tenfold increase in natural killer activity and a three- to four-fold increase in anti-CD3-induced IFN-γ production in cultured splenocytes. Furthermore, when challenged with Renca tumor cells at a distant site, the day of tumor emergence was significantly delayed, and tumor growth was reduced in mice that received IL-12 cDNA, compared to mice given injections of plasmid vector alone. A number of the mice receiving IL-12 cDNA injections remained tumor free months after tumor challenge (26). Responses follow the same pattern when mice are challenged with MethA, SA-1, L5178 and P815 tumor cells (27).

Clinical trials of DNA vaccines

Prophylactic and therapeutic DNA vaccine clinical trials for a variety of pathogens and cancers are underway (Table 2). All candidate vaccines are in early-stage trials examining safety and immune responses.

Humoral and cellular immune responses have been shown to be induced by DNA vaccines expressing HIV-1 genes, suggesting prophylactic and therapeutic value for the plasmids. Vaccination with these plasmids has decreased HIV-1 viral load in infected chimpanzees. Furthermore, immunized chimpanzees were protected against a challenge with HIV-1. Clinical trials for a HIV gp160 DNA-based vaccine are underway. Phase I trial with the vaccine induced a good safety profile and demonstrated an immunological potentiation (28).

Table 2. Human trials of DNA vaccines

Vaccines against	Proteins encoded by DNA vaccines	Results	
		antibodies	CTL
HIV	Envelope and regulatory proteins, or core proteins and enzyme involved in HIV replication	ND	+
Hepatitis B virus	HBs	+	+
Herpes simplex virus	Herpes glycoprotein	UE	UE
Influenza virus	HA	UE	UE
Malaria (*Plasmodium sp*)	CSP, Spf66	ND	+
Adenocarcinoma (colon and breast)	CEA	ND	+
B cell lymphoma	Immunoglobulin	+	ND
Cutaneous T cell lymphoma	T cell receptor	UE	UE
Prostate cancer	Prostate-specific membrane antigen	UE	UE

ND, not determined; UE, under examination.

HBs antigens were delivered by the PowderJect XR1 gene delivery system into human skin. Only one of seven seronegative volunteers developed high titers of HBs antibody. The volunteer may have had previous exposure to hepatitis B, suggesting that HBs DNA vaccine given by the gene delivery system may induce a booster response (29).

Malaria caused by *Plasmodium falciparum* is the most important parasitic disease and difficult to control. Spf66 is a synthetic polypeptide based on pre-erythrocytic and asexual blood-stage proteins of *P. Falciparum*. Clinical phase III trials of this vaccine undertaken in Laten America and in Africa have documented partial efficacy against clinical malaria. To improve the vaccine efficacy, research efforts with DNA vaccine encoding Spf66 are now ongoing (30).

Since CTL play a pivotal role in rejection of tumor cells, DNA vaccine trials are applied to induce anti-tumor immunity. As colorectal cancer antigens such as 17-1A, 791Tgp72 and caricinoembryonic antigen (CEA) have been identified, relevant articles were published on DNA vaccines for colorectal cancer. A number of approaches are currently being evaluated in Phase I, II and III trials (31). Genetically modified tumor vaccines have been also employed in clinical trials. Immunization with autologous tumor cells transfected with the GM-CSF gene or the B7 gene produced specific immune responses and objective clinical responses with minimal toxicity in phase I/II trials (32).

Since an immunostimulatory polynucleotide DNA sequence (ISS) containing the CpG motif induces the Th1-type response, the ISS is being employed for therapy of allergic diseases (type I hypersensitivity). The primary basis for allergy is an imbalance in the immune response resulting in production of Th2 cytokines (IL-4 and IL-5), and antibodies of the IgE isotype, that promote the symptoms of allergic disease. The aim of using ISS is to "re-balance" this response and alter the underlying mechanism of disease. The Phase I open-label study compared the skin test responses of 6 ragweed allergic subjects to a conjugate of purified ragweed allergen (Amb a 1) to an ISS versus a licensed ragweed extract. The skin test allergic response to the conjugate was significantly less than that of the licensed ragweed extract.

Conclusion

The initial successes of DNA vaccines in generating protective immunity against viral infection have been repeated with vaccines against bacterial infection and cancer. DNA vaccines could potentially take the place of live attenuated vaccines, which have been required for generation of cellular immune responses against intracellular pathogens. Several DNA vaccines have now entered phase I/II human clinical trials. There are several hurdles to be overcome on the road to the use of DNA vaccine clinically. These include improving gene delivery and potency so that low doses of DNA can achieve the efficacy of conventional vaccines.

REFERENCES

1 Davis HL, Schirmbeck R, Reimann J and Whalen RG: DNA-mediated immunization in mice induces a potent MHC class I-restricted cytotoxic T lymphocyte response to the hepatitis B envelope protein. Hum Gene Ther **6**, 1447 - 1456 (1995)

2 Raz E, Tighe H, Sato Y, Corr M, Dudler JA, Roman M, Swain SL, Spiegelberg HL and Carson DA: Preferential induction of a Th1 immune response and inhibition of specific IgE antibody formation by plasmid DNA immunization. Proc Natl Acad Sci USA **93**, 5141 - 5145 (1996)

3 Ying H, Zaks TZ, Wang RF, Irvine KR, Kammula US, Marincola FM, Leitner WW and Restifo NP: Cancer therapy using a self-replicating RNA vaccine. Nat Med **5**, 823 - 827 (1999)

4 Davis HL, Michel ML and Whalen RG: DNA-based immunization induces continuous secretion of hepatitis B surface antigen and high levels of circulating antibody. Hum Mol Genet **2**, 1847 - 1851 (1993)

5 Danko I, Fritz JD, Jiao S, Hogan K, Latendresse JS and Wolff JA: Pharmacological enhancement of in vivo foreign gene expression in muscle. Gene Ther **1**, 114 - 121 (1994)

6 Pertmer TM, Roberts TR and Haynes JR: Influenza virus nucleoprotein specific immunoglobulin G subclass and cytokine responses elicited by DNA vaccination are dependent on the route of vector DNA delivery. J Virol **70**, 6119 - 6125 (1996)

7 Corr M, Lee DJ, Carson DA and Tighe H: Gene vaccination with naked plasmid DNA: mechanism of CTL priming. J Exp Med **184**, 1555 - 1560 (1996)

8 Robinson HL and Torres CA: DNA vaccines. Semin Immunol **9**, 271 - 283 (1997)

9 Torres CA, Iwasaki A, Barber BH and Robinson HL: Differential dependence on target site tissue for gene gun and intramuscular DNA immunizations. J Immunol **158**, 4529 - 4532 (1997)

10 Feltquate DM, Heaney S, Webster RG and Robinson HL: Different T helper cell types and antibody isotypes generated by saline and gene gun DNA immunization. J Immunol **158**, 2278 - 2284 (1997)

11 Krieg AM, Yi AK, Matson S, Waldschmidt TJ, Bishop GA, Teasdale R, Koretzky GA and Klinman DM: CpG motifs in bacterial DNA trigger direct B cell activation. Nature **374**, 546 - 549 (1995)

12 Klinman DM, Yi AK, Beaucage SL, Conover J and Krieg AM: CpG motifs present in bacteria DNA rapidly induce lymphocytes to secrete interleukin 6, interleukin 12, and interferon gamma. Proc Natl Acad Sci USA **93**, 2879 - 2883 (1996)

13 Yi AK, Tuetken R, Redford T, Waldschmidt M, Kirsch J and Krieg AM: CpG motifs in bacterial DNA activate leukocytes through the pH- dependent generation of reactive oxygen species. J Immunol **160**, 4755 - 4761 (1998)

14 Yoshida A, Nagata T, Uchijima M, Higashi T and Koide Y: Advantage of gene gun-mediated over intramuscular inoculation of plasmid DNA vaccine in reproducible induction of specific immune responses. Vaccine **18**, 1725 - 1729 (2000)

15 Ulmer JB, Deck RR, De Witt CM, Friedman A, Donnelly JJ and Liu MA: Protective immunity by intramuscular injection of low

doses of influenza virus DNA vaccines. Vaccine **12**, 1541 - 1544 (1994)

16 Tsuji T, Hamajima K, Fukushima J, Xin KQ, Ishii N, Aoki I, Ishigatsubo Y, Tani K, Kawamoto S, Nitta Y, Miyazaki J, Koff WC, Okubo T and Okuda K: Enhancement of cell-mediated immunity against HIV-1 induced by coinnoculation of plasmid-encoded HIV-1 antigen with plasmid expressing IL-12. J Immunol **158**, 4008 - 4013 (1997)

17 Koide Y, Nagata T, Yoshida A and Uchijima M: DNA vaccines for infections with intracellular bacteria. Curr Trends Immunol **1**, 123 - 132 (1998)

18 Uchijima M, Yoshida A, Nagata T and Koide Y: Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I restricted T cell responses against an intracellular bacterium. J Immunol **161**, 5594 - 5599 (1998)

19 Nagata T, Uchijima M, Yoshida A, Kawashima M and Koide Y: Codon optimization effect on translational efficiency of DNA vaccine in mammalian cells: analysis of plasmid DNA encoding a CTL epitope derived from microorganisms. Biochem Biophys Res Commun **261**, 445 - 451 (1999)

20 Tascon RE, Colston MJ, Ragno S, Stavropoulos E, Gregory D and Lowrie DB: Vaccination against tuberculosis by DNA injection. Nat Med **2**, 888 - 892 (1996)

21 Huygen K, Content J, Denis O, Montgomery DL, Yawman AM, Deck RR, De Witt CM, Orme IM, Baldwin S, D'Souza C, Drowart A, Lozes E, Vandenbussche P, Van Vooren JP, Liu MA and Ulmer JB: Immunogenicity and protective efficacy of a tuberculosis DNA vaccine. Nat Med **2**, 893 - 898 (1996)

22 Lozes E, Huygen K, Content J, Denis O, Montgomery DL, Yawman AM, Vandenbussche P, Van Vooren JP, Drowart A, Ulmer JB and Liu MA: Immunogenicity and efficacy of a tuberculosis DNA vaccine encoding the components of the secreted antigen 85 complex. Vaccine **15**, 830 - 833 (1997)

23 Montgomery DL, Huygen K, Yawman AM, Deck RR, Dewitt CM, Content J, Liu MA and Ulmer JB: Induction of humoral and cellular immune responses by vaccination with *M. tuberculosis* antigen 85 DNA. Cell Mol Biol (Noisy-le grand) **43**, 285 - 292 (1997)

24 Lowrie DB, Tascon RE, Bonato VL, Lima VM, Faccioli LH, Stavropoulos E, Colston MJ, Hewinson RG, Moelling K and Silva CL: Therapy of tuberculosis in mice by DNA vaccination. Nature **400**, 269 - 271 (1999)

25 Bueler H and Mulligan RC: Induction of antigen-specific tumor immunity by genetic and cellular vaccines against MAGE: enhanced tumor protection by coexpression of granulocyte-macrophage colony-stimulating factor and B7-1. Mol Med **2**, 545 - 555 (1996)

26 Tan J, Newton CA, Djeu JY, Gutsch DE, Chang AE, Yang NS, Klein TW and Hua Y: Injection of complementary DNA encoding interleukin-12 inhibits tumor establishment at a distant site in a murine renal carcinoma model. Cancer Res **56**, 3399 - 3403 (1996)

27 Rakhmilevich AL, Turner J, Ford MJ, McCabe D, Sun WH, Sondel PM, Grota K and Yang NS: Gene gun-mediated skin transfection with interleukin 12 gene results in regression of established primary and metastatic murine tumors. Proc Natl Acad Sci USA **93**, 6291 - 6296 (1996)

28 Ugen KE, Nyland SB, Boyer JD, Vidal C, Lera L, Rasheid S, Chattergoon M, Bagarazzi ML, Ciccarelli R, Higgins T, Baine Y, Ginsberg R, Macgregor RR and Weiner DB: DNA vaccination with HIV-1 expressing constructs elicits immune responses in humans. Vaccine **16**, 1818 - 1821 (1998)

29 Tacket CO, Roy MJ, Widera G, Swain WF, Broome S and Edelman R: Phase 1 safety and immune response studies of a DNA vaccine encoding hepatitis B surface antigen delivered by a gene delivery device. Vaccine **17**, 2826 - 2829 (1999)

30 Tanner M and Alonso PL: The development of malaria vaccines: SPf66- what next? Schweiz Med Wochenschr **126**, 1210 - 1215 (1996)

31 Maxwell-Armstrong CA, Durrant LG and Scholefield JH: Colorectal cancer vaccines. Br J Surg **85**, 149 - 154 (1998)

32 Mahvi DM, Sondel PM, Yang NS, Albertini MR, Schiller JH, Hank J, Heiner J, Gan J, Swain W and Logrono R: Phase I/IB study of immunization with autologous tumor cells transfected with the GM-CSF gene by particle mediated transfer in patients with melanoma or sarcoma. Hum Gene Ther **8**, 875 - 891 (1997)

Codon usage limitation in the expression of HIV-1 envelope glycoprotein

Jürgen Haas*‡, Eun-Chung Park† and Brian Seed*†

Background: The expression of both the *env* and *gag* gene products of human immunodeficiency virus type 1 (HIV-1) is known to be limited by *cis* elements in the viral RNA that impede egress from the nucleus and reduce the efficiency of translation. Identifying these elements has proven difficult, as they appear to be disseminated throughout the viral genome.

Results: Here, we report that selective codon usage appears to account for a substantial fraction of the inefficiency of viral protein synthesis, independent of any effect on improved nuclear export. The codon usage effect is not specific to transcripts of HIV-1 origin. Re-engineering the coding sequence of a model protein (Thy-1) with the most prevalent HIV-1 codons significantly impairs Thy-1 expression, whereas altering the coding sequence of the jellyfish green fluorescent protein gene to conform to the favored codons of highly expressed human proteins results in a substantial increase in expression efficiency.

Conclusions: Codon-usage effects are a major impediment to the efficient expression of HIV-1 genes. Although mammalian genes do not show as profound a bias as do *Escherichia coli* genes, other proteins that are poorly expressed in mammalian cells can benefit from codon re-engineering.

Addresses: *Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. †Melvin and Barbara Nessel Gene Therapy Center, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.

‡Present address: Institut für Medizische Virologie, Universität Heidelberg, Im Neuenheimer Feld 324, 69120 Heidelberg, Germany.

Correspondence to: Brian Seed
E-mail address: seed@molbio.mgh.harvard.edu

Received: **21 December 1995**
Revised: **31 January 1996**
Accepted: **31 January 1996**

Current Biology 1996, Vol 6 No 3:315–324

 ISSN 0960-9822

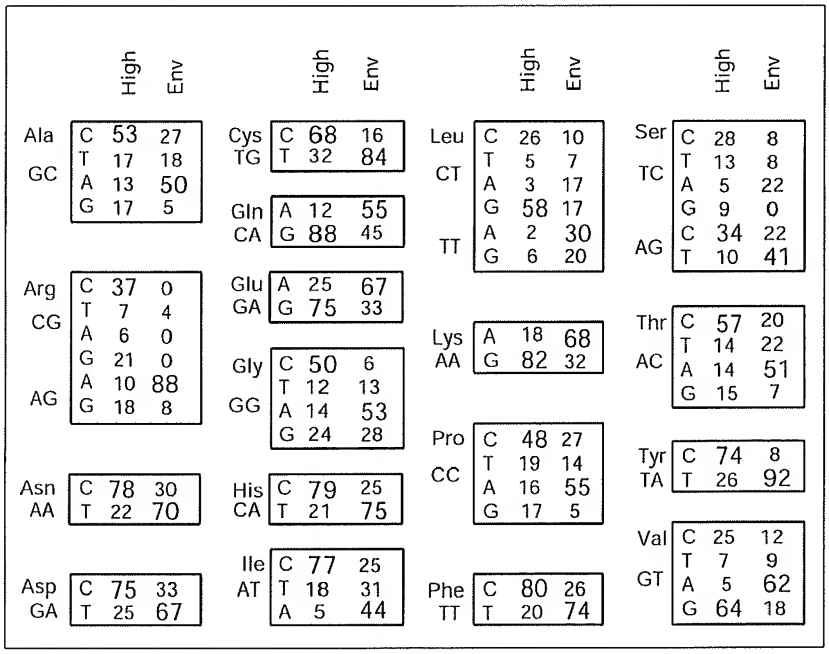
Background

The mature envelope glycoproteins of human immunodeficiency virus type 1 (HIV-1), gp120 and gp41, form a heterodimeric membrane complex embedded in the lipid bilayer that circumscribes the viral inner capsid. Expression of envelope proteins, like those of the gag polyprotein and reverse transcriptase, is facilitated by a *cis*-acting element in the viral mRNA known as the rev-responsive element or RRE [1–7], and by the action of rev, a small RNA-binding polypeptide encoded by a subgenomic RNA overlapping the coding region of the envelope glycoproteins [8,9].

Rev has been reported to exert its effect at two or more stages in the expression of a viral gene product: first by expediting the export to the cytoplasm of viral RNAs bearing the RRE [1,4,5,10–16], and subsequently by improving the translational potency of RRE-containing transcripts [17–19]. What causes viral mRNAs to be inadequately expressed is not clear. Several studies point to the existence of poorly localized sequences within the *gag* and *env* coding regions that attenuate expression [10,20–24]. One of these inhibitory sequences has been shown to overlap the RRE itself [25–27], but others have proven more difficult to localize. For example, the concerted change of four coding segments, of approximately 30 nucleotides each, within the 5′ end of the *gag* coding sequence produced a transcript that was substantially *rev*-independent; further dissection of the inhibitory sequences proved difficult, however, as mutation of all four regions was required to sustain the increased protein synthesis [28]. Cellular factors, including heterogenous ribonucleoprotein C (hnRNP C), or a serologically related protein, have been found to form complexes with a 270 residue inhibitory sequence, suggesting that intranuclear retention may be mediated by sequence-specific interactions [29]. Nonetheless, an explanation for the lower translational efficiency of the mRNA template remains to be advanced.

The envelope proteins of HIV-1 are a natural target for vaccines and for post-infection treatments to limit viral spread. Despite a great deal of effort aimed at developing and studying the viral envelope proteins, little is known about the factors that limit their expression, the role of non-CD4 cofactors in viral entry, or the mechanisms of tissue tropism. The poor expression of envelope proteins limits the titer of retroviral pseudotypes, making it difficult to develop analytical tools that restrict the success or failure of infection to membrane-proximal events. In attempting to address this problem, we considered the profound bias in the codon usage of the *env* proteins, a bias that extends to the *gag* and *pol* proteins [30–33]. Figure 1 illustrates the divergence between the codon prevalence found in the coding region for the envelope protein of the HIV-1 LAV isolate and that found in a compendium of highly expressed human genes. Here, we report on the consequences of codon bias, and on the favorable outcome

Figure 1



Codon usage of HIV-1 envelope (env) and highly expressed human (high) genes. The frequencies (×100) of the individual codons are shown for each of the degenerately encoded amino acids, and the most prevalent codon is shown in bold.

of the systematic replacement of the native codons of gp120 with codons chosen to reflect more closely the codon preference of highly expressed human genes.

The surprising efficacy of codon replacement for gp120 led us to explore the consequences of codon re-engineering for the expression of two small test proteins: Thy-1, an abundant cell-surface protein of rodent thymocytes, and green fluorescent protein (GFP) [34–36], a jellyfish (*Aequorea victoria*) protein that is relatively poorly expressed in mammalian cells. Conversion of the Thy-1 coding sequence to that of a synthetic gene with the codon preference of HIV-1 envelope protein gave severely attenuated expression, whereas replacement of the endogenous codons of GFP with those of highly expressed mammalian proteins resulted in a substantial increase in synthetic efficiency.

Results

Construction of a synthetic gp120 gene based on optimal codon usage

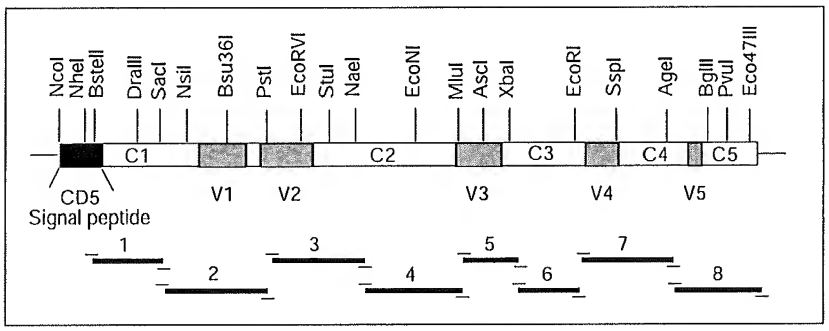
To explore whether codon bias accounted for the poor expression of envelope glycoproteins we constructed a synthetic gene encoding the gp120 segment of HIV-1, based on the sequence of the prototype virus of the most common North American subtype, HIV-1MN. The synthetic gp120 was assembled from chemically synthesized long oligonucleotides that were subsequently amplified in the crude state by polymerase chain reaction (PCR). Some deviations from strict adherence to optimized codon usage were made to accommodate the introduction of unique restriction sites into the resulting gene at approximately 100 base-pair intervals (Fig. 2). In addition, because the endogenous secretory peptide of envelope is known to function inefficiently [37], it was replaced by the leader peptide of the human CD5 antigen, which efficiently directs the synthesis and export of secreted and membrane-bound proteins [38]. The chimeric gp120 precursor gene was inserted in a mammalian cDNA expression vector under the control of the human cytomegalovirus (CMV) immediate early promoter.

Expression of wild-type and synthetic gp120

To evaluate the relative potency of the wild-type and synthetic gp120 coding sequences, we compared the outcomes of transient transfection of the synthetic and native genes, as well as the result of replacement of the endogenous gp120 secretory leader sequence with that of the human CD5 antigen. As shown in Figure 3, the synthetic gene product was expressed at a very high level compared with that of the wild-type gp120, whether expressed with the native or CD5 leader, and as assessed either by immunoprecipitation with a CD4–immunoglobulin fusion protein or by enzyme-linked immunosorbent assay; in the latter assay, expression of synthetic gp120 exceeded that of the native protein with CD5 leader by 41-fold to several hundred-fold, depending on the experiment. Because optimal expression of native envelope sequences requires both the RRE in *cis* and *rev* in *trans*, a similar experiment was conducted in which both the synthetic and endogenous genes were endowed with the RRE, and expression

Figure 2

Synthesis of a gp120 (HIV-1MN) gene with the codon preference of highly expressed human genes. The shaded portions marked V1 to V5 indicate segments of the envelope that show high variability between different natural isolates, and C1 through C5 denote regions that remain relatively constant by the same criteria. Unique restriction sites engineered into the sequence are shown above, and a bar diagram of the chemically synthesized DNA fragments which served as PCR templates is shown below the sequence.



was measured by immunoprecipitation following cotransfection of the envelope expression plasmids and a plasmid expressing *rev*. The results show that *rev* had no effect on the expression of the synthetic gene product, but significantly enhanced expression of the native gene product (Fig. 4). Thus, the action of *rev* is not apparent on a substrate which lacks the coding sequence of the endogenous viral envelope sequences. Because *rev* appears to exert its effect at two steps in the expression of a viral transcript, we sought to clarify the possible role of export in the improved expression of the synthetic gene; we therefore turned to expression from recombinant vaccinia viruses.

Cytoplasmically transcribed synthetic gp120 is more efficiently expressed

Native gp120 contains a transcriptional termination sequence for vaccinia early transcripts, so we used for comparison a variant virus from which this sequence had been removed for higher production [39]. As shown in Figure 5, increased expression of the synthetic gene was demonstrable when the endogenous gene product and the synthetic gene product were expressed from vaccinia virus recombinants under the control of the strong mixed early and late 7.5 k promoter. Densitometry showed the improvement to be from 8–50 fold, depending on the cell type. As vaccinia virus transcripts are created and translated in the cytoplasm of infected cells, the increased expression of the synthetic gp120 gene in this circumstance cannot be attributed to improved export from the nucleus. RNA blot analysis showed that the vaccinia transcripts encoding the synthetic gp120 were less abundant than the transcripts encoding native gp120, indicating that the increased efficiency of expression was not attributable to increased mRNA stability (Fig. 6).

Figure 3

Expression of synthetic gp120 in transient transfection assays. **(a)** Gel electrophoresis of immunoprecipitated supernatants of 293T cells transfected with plasmids expressing gp120 encoded by the IIIB isolate of HIV-1, (gp120IIIb), the MN isolate (gp120mn), the MN isolate modified by substitution of the endogenous leader peptide with that of the CD5 antigen (gp120mnCD5L), or the chemically synthesized gene encoding the MN variant with the human CD5 leader (syngp120mn). Supernatants were harvested following a 12 h labeling period, 40 h post-transfection, and immunoprecipitated with CD4–IgG1 fusion protein and protein A agarose. **(b)** ELISA of supernatants of transiently transfected cells. Supernatants of 293T cells transfected by calcium phosphate with the constructs described above were harvested after 4 days and tested in a gp120/CD4 ELISA.

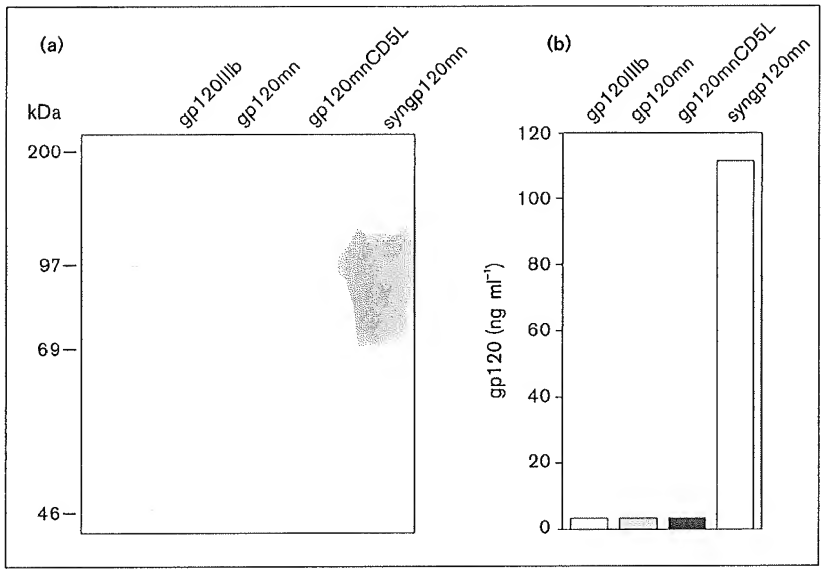
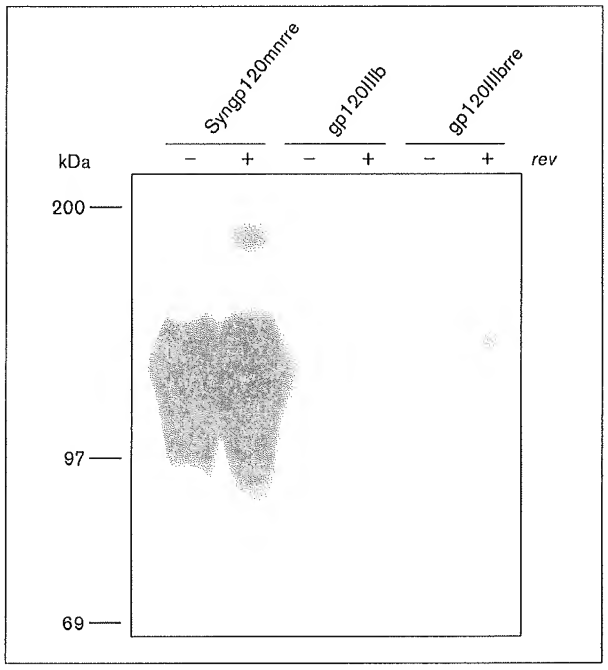
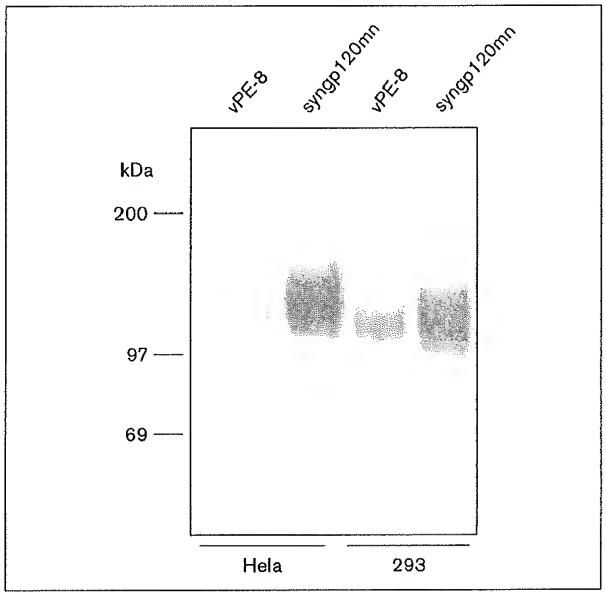


Figure 4



Expression of synthetic and wild-type envelope sequences in the presence of *rev* in *trans* and the RRE in *cis*. 293T cells were transiently transfected by calcium phosphate coprecipitation of 10 μg plasmid expressing: the synthetic gp120MN sequence and RRE in *cis* (syngp120mnrre), the gp120 portion of HIV-1 IIIB (gp120IIIb), and the same sequence with the RRE in *cis* (gp120IIIbrre); each gp120 expression plasmid was cotransfected with 10 μg of either pCMVrev or CDM7 plasmid DNA. Supernatants were harvested 60 h post-transfection and immunoprecipitated with CD4–IgG fusion protein and protein A agarose. The gel exposure time was extended to allow the induction of gp120IIIBrre by *rev* to be demonstrated. Shorter exposures showed no significant dependence of syngp120mnrre expression upon *rev*.

Figure 5



Cytoplasmic expression of syngp120mn by vaccinia virus. Immunoprecipitation of supernatants of human 293 or Hela cells infected with vaccinia virus expressing wild-type gp120IIIb (vPE-8) or synthetic gp120mn (syngp120mn). Cells were infected at a multiplicity of infection of at least 10. Supernatants were harvested after 24 h of infection and immunoprecipitated with CD4–IgG fusion protein and protein A agarose.

HIV codon patterns undermine the efficiency of Thy-1 expression

To explore further the importance of codon usage in envelope protein expression, we replaced the codons of a small, typically highly expressed cell-surface protein, the rat Thy-1 antigen, with the codons most frequently used by the HIV-1 envelope protein. The resulting sequence was then edited to remove any message destabilization motifs of the form AUUUA that had been created, and to introduce two restriction sites for ease of creation and manipulation of the resulting sequence. The synthesis strategy is shown in Figure 7. As shown in Figure 8, the synthetic rat Thy-1 was expressed approximately 100-fold less well than the native gene.

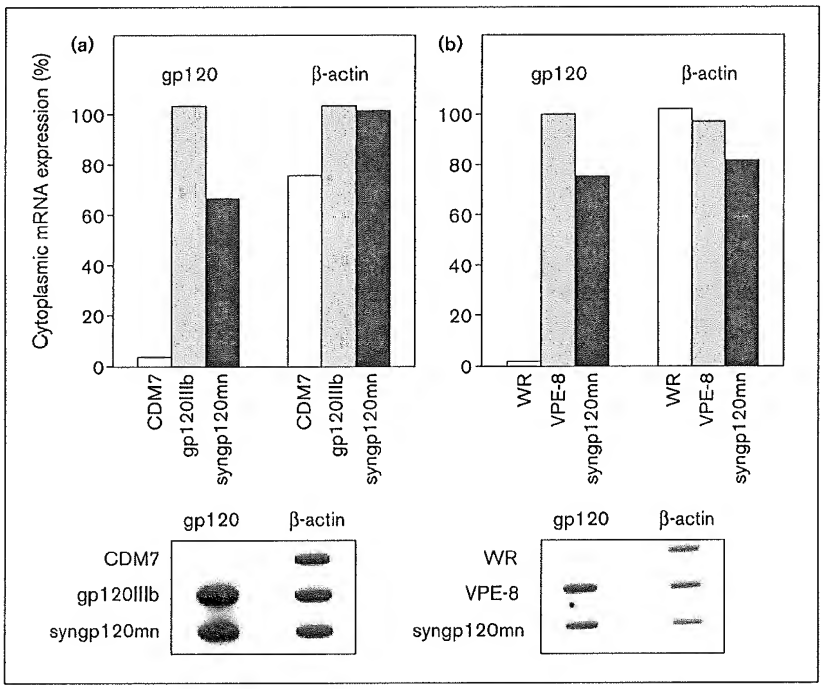
As codon usage for highly expressed proteins can vary from organism to organism, sequence re-engineering may prove to be a generally useful way to improve the expression of genes poorly expressed in mammalian cells. To explore the generality of this approach we selected a second small model protein.

Codon replacement enhances green fluorescent protein expression

The GFP of the jellyfish *Aequorea victoria* [34–36] has attracted attention recently for its possible use as a marker or reporter for transfection and lineage studies [40], but has been found to be expressed poorly in mammalian cells. Examination of a codon usage table constructed from the native coding sequence showed that the GFP codons favored either A or U in the third position. The bias in this case favors A less than does the bias of gp120, but is substantial. A synthetic gene was created in which the GFP sequence was recreated in the same manner as for gp120, and the initiation consensus was replaced with sequences corresponding to the translational initiation consensus. The expression of the resulting protein was contrasted with that of the wild-type sequence, similarly engineered to bear an optimized translational initiation consensus (Fig. 9). In addition, the effect of replacing the serine at position 65 with a threonine (S65T), reported to improve excitation efficiency at 490 nm and hence preferred for fluorescence microscopy [41], was examined (Fig. 9).

Figure 6

Increased mRNA concentration does not account for the higher output of synthetic gp120. Cytoplasmic RNA was prepared from (a) COS cells or (b) 293 cells transfected with expression plasmids (a) or infected with recombinant vaccinia viruses (b). RNA was quantitated by slot blot and the relative concentrations are shown above images of the slot blot autoradiogram. Actin concentrations are shown as a control. In all cases the mRNA concentration of the synthetic gene was less than that of the native gene. CDM7, expression vector only. WR, wild-type vaccinia. vPE-8, recombinant vaccinia expressing native gp120IIIb.



Codon engineering conferred a significant increase in expression efficiency (and concomitant percentage of cells apparently positive for transfection), and the combination of the S65T mutation and codon optimization resulted in a DNA segment encoding a highly visible mammalian marker protein (Fig. 9). There was a net improvement in fluorescence per cell of between 40–120 fold, depending on detection conditions.

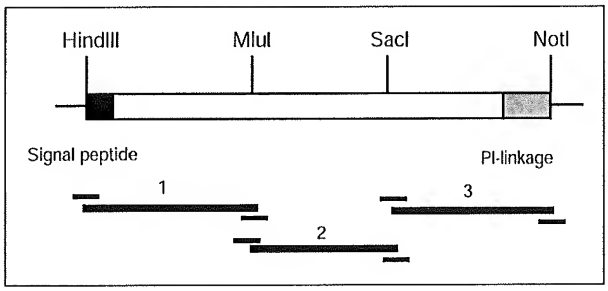
Discussion

Phenomenological rules for codon patterns

Examination of Figure 1 shows that three simple rules describe the pattern of envelope codon usage: first, preferred codons maximize the number of adenine residues in the viral RNA; second, T is preferred over C whenever the third position degeneracy is a pyrimidine; and third, the dinucleotide CG is highly under-represented. In all cases but one, the first rule implies that, if the third position can be A, that codon will be most frequently used [31]. In the special case of serine, three codons equally contribute one A residue to the mRNA; together these three comprise 85 % of the codons found in envelope transcripts. A particularly striking example of the combined effects of first and third rules is found in the codon choice for arginine, in which the AGA triplet comprises 88 % of all codons. Finally, the third rule implies that the third position is much less likely to be G whenever the second position is C, as in the codons for alanine, proline, serine and threonine (Fig. 1).

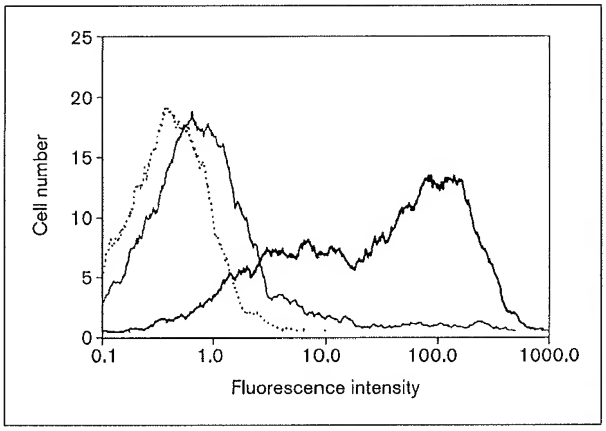
Several studies have pointed to a tendency for HIV reverse transcriptase to exhibit a high frequency of G to A transition, a phenomenon called G to A hypermutation [42–44], and it is known that other lentiviruses undergo similar hypermutation [45]. Although this helps to account for the general prevalence of A and T, additional hypotheses are needed to account for the bias of A over T in the third position, a bias that appears to reflect a more general excess of purines in the viral RNA strand. One suggestion is that the nucleotide bias should be restated as high A, low C, and is driven by two factors: the low intracellular

Figure 7



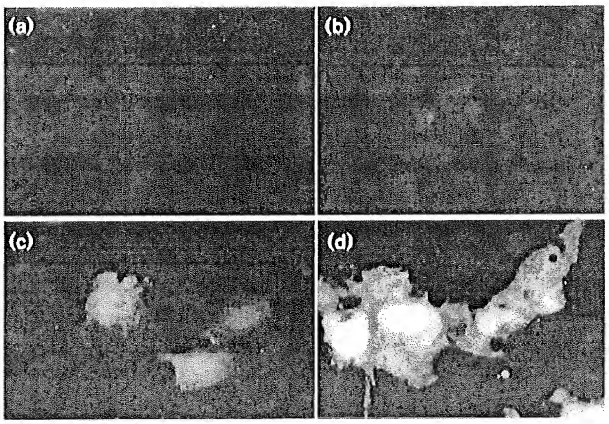
Schematic diagram of a rat Thy-1 gene with HIV-1 envelope codon usage (rTHY-1env). The darkened boxes in the rTHY-1env construct denote the leader peptide and the sequences in the precursor that direct the attachment of a phosphatidylinositol glycan anchor. Unique restriction sites used for assembly of the construct are shown above.

Figure 8



Surface expression of rat Thy-1 with HIV-1 envelope codon usage. Shown are flow cytometry histograms of 293T cells transiently transfected with either wild-type rat Thy-1 (thick line), rat Thy-1 with envelope codons (thin line) or vector only (dotted line). 293T cells were transfected with the different expression plasmids by calcium phosphate coprecipitation and stained with anti-rat Thy-1 monoclonal antibody OX7 followed by a polyclonal FITC-conjugated anti-mouse IgG antibody 3 days after transfection.

Figure 9



Expression of GFP. (a) COS cells transfected with vector only. (b) COS cells transfected with a CDM7 expression plasmid encoding native GFP engineered to include a consensus translational initiation sequence. (c) COS cells transfected with an expression plasmid having the same flanking sequences and initiation consensus as in (b), but bearing a codon-optimized gene sequence. (d) COS cells transfected with an expression plasmid as in (c), but encoding GFP(S65T).

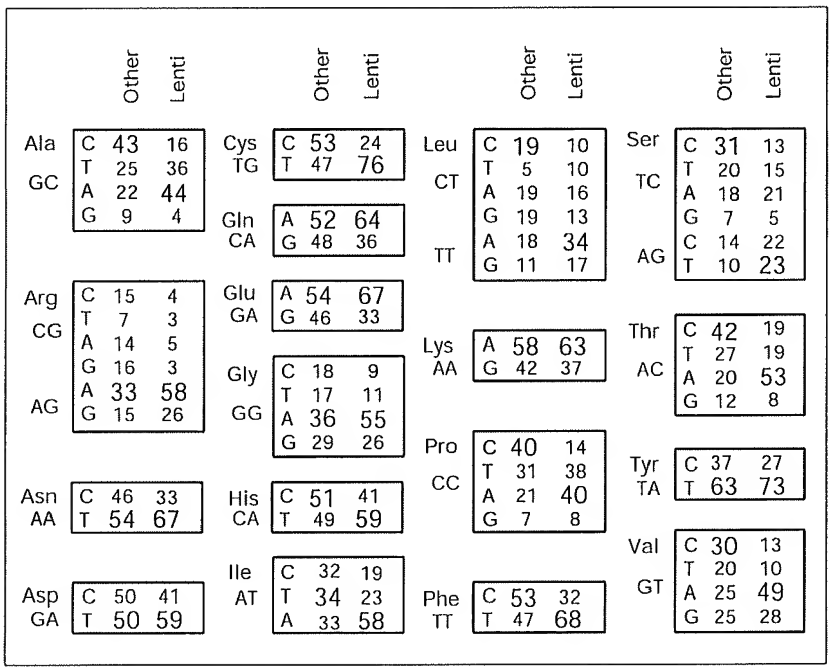
concentration of dCTP relative to TTP; and the higher error rate of HIV reverse transcriptase on RNA templates than on DNA, which may promote transversions from pyrimidine residues in the RNA ([44,46]; S. Wain-Hobson, personal communication).

Why should inefficient codons be prevalent in HIV-1?

Although it is often thought that viruses occupy an evolutionary niche which places a premium on rapid replication, persistent infection of a fraction of susceptible individuals is a common feature of several viral strains and may be important for the maintenance of a reservoir of viral genomes that can be transmitted to new hosts. This appears to be especially true of HIV-1 and HIV-2. Codon bias may allow the expression of such viruses to be suppressed in order to minimize the antigenic profile. An opposing hypothesis, that the presence of adenine residues in genomic RNA is generally favored in the retroviral life cycle — for example, by facilitating the process of replication — is not consistent with earlier reports that Visna virus and HIV-1 share similar codon usage, whereas human T-cell leukemia virus type 1 (HTLV-1) and HIV-1 do not [30], or that murine leukemia virus and HIV-1 are similarly discordant [31]. Figure 10 illustrates this point with a compilation of codon frequencies chosen from envelope glycoproteins of a variety of other retroviruses — but excluding the lentiviruses — compared with those drawn from the envelope sequences of five lentiviruses that are not closely related to HIV-1: bovine immunodeficiency virus, caprine arthritis encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and Visna virus. The codon usage patterns for the lentiviruses is strikingly similar to that of HIV-1 — in all cases the preferred codon for HIV-1 is the same as the preferred codon for the other lentiviruses. In contrast, the non-lentiviral envelope codons do not show a similar predominance of A residues, and are also not as skewed toward third position C and G residues as are the highly expressed human genes. In general the non-lentiviral retroviruses appear to exploit the different codons more equally, a pattern they share with less highly expressed human genes (data not shown). In addition to the prevalence of codons including A, the lentiviral codons also show the HIV-1 pattern of strong CpG under-representation, such that the third position for alanine, proline, serine and threonine triplets is rarely G. The retroviral envelope triplets of the comparison group show a similar, but less pronounced under-representation of CpG. The most obvious difference between lentiviruses and other retroviruses with respect to CpG prevalence lies in the usage of the CGX variant of arginine triplets, which is relatively frequently represented among the retroviral envelope coding sequences, but is almost never present among lentiviral sequences. The under-representation of CpG-bearing triplets in lentiviruses could be a consequence of genetic selection to avoid transcriptional silencing by methylation of CpG cytosines [47], consistent with the finding that CpG dinucleotides are also under-represented among untranslated segments of the HIV-1 genome (data not shown).

Figure 10

Codon usage of lentiviral (lenti) and other retroviral envelope genes. The frequencies (×100) of the individual codons are shown for each of the degenerately encoded amino acids, and the most prevalent codon is shown in bold.



Representatives of two unusual types of retrovirus were not included in the group compared with the lentiviruses in Figure 10: the spumaretroviruses, exemplified by bovine syncytial virus and the human and simian spumaviruses; and Walleye dermal sarcoma virus (WDSV), which appears to constitute a new genus [48]. Viruses of these two classes show envelope codon-usage patterns similar to those of the lentiviruses and, like the lentiviruses, have large genomes that encode several short open reading frames in addition to the *gag*, *pol* and *env* constituents [48]. Intriguingly, the spumaviruses, WDSV, and at least some of the lentiviruses share the presence of single-stranded gaps in their unintegrated linear DNA [48–52]. This small constellation of similar attributes suggests they may have arisen from a common progenitor.

Why is there codon preference in any form?

Codon bias has been observed in many species [53–56], and a correlation has been noted between high expression and the use of a stereotyped pattern of codons [54,57]. Although this correlation is not universal ([58]; see also discussion in [59]), in several cases it has been found that expression of exogenous gene products in *Escherichia coli* can be enhanced by systematic substitution of the endogenous codons with triplets over-represented in highly expressed *E. coli* genes [59,60]. Although highly expressed mammalian genes show nonrandom codon-usage patterns, the degree of over-representation of favored codons is not as pronounced as for highly expressed bacterial genes. Nonetheless, the data presented here clearly suggest that codon usage can play an important role in determining translational efficiency in a mammalian cell context.

It has been widely assumed that translational efficiencies of mammalian gene products are governed by initiation; if they were not, mRNAs would, in general, be maximally loaded with ribosomes. This study suggests that poorly translated mRNAs may indeed be maximally loaded. However, the finding that inferior codons limit translational efficiency is not necessarily inconsistent with the view that translational efficiency is governed by initiation. First, the limitation imposed by inferior codons need not be kinetic. Instead, an abundance of less favored codons could incur a significant cumulative probability of failure to complete the nascent polypeptide. Recent evidence, however, suggests that premature termination does not occur with a high enough efficiency to account for the observed stimulation (E.C.P. and B.S., unpublished observations). Another possibility is that the inferior codons reflect a selection for RNA structures that influence the rate of initiation — for example, if access to ribosomes is controlled by cues distributed throughout the RNA. In such a case, lentiviral codons could predispose the RNA to accumulate in a pool of poorly initiated RNAs. The sequestered RNA might be given an improved rate of initiation by the action of *rev* or other regulatory proteins. No stimulation by *rev* would be seen in the event that the

RNA did not bear the unfavored sequences, consistent with these results. Some precedent for the existence of discrete pools of mRNA can be found in studies of maternal transcript distribution and use in early embryogenesis [61,62]. RNA adenine methylation, a candidate RNA modification that might dictate transcript use, seems unlikely to participate in translational suppression because the known methylation consensus [63] is not substantially over-represented among HIV sequences relative to other retroviruses (data not shown; there is a profound strand bias in the frequency of the RNA methylation consensus, however, which is consistent with the purine excess in the sense strand).

Conclusions

The creation of a synthetic coding sequence based on codons over-represented in highly expressed human genes overcomes a major limitation to the translational efficiency of HIV-1 envelope glycoprotein. Although codon optimization has been used in the past to improve expression in *E. coli* of genes from other organisms, similar studies have not been conducted in mammalian cells, perhaps because the codon bias of highly expressed mammalian genes is not as striking as that of *E. coli* genes. However, the results obtained here with three unrelated proteins — HIV-1 gp120, rat cell-surface antigen Thy-1 and GFP from *Aequorea victoria* — suggest that codon optimization may prove to be a fruitful strategy for improving the expression in mammalian cells of genes that show limited translational efficiency in their native form.

Materials and methods

Plasmid constructions

The synthetic gp120 gene was generated from eight 160–200 base oligonucleotides, produced on a Milligen 8750 synthesizer. After elution with 30 % ammonium hydroxide, the oligonucleotides were deblocked at 55 °C for 12 h, precipitated with n-butanol and resuspended in H_2O. 15–25-mer oligonucleotides complementary to the ends were used to amplify the long oligonucleotides by PCR. Typically, PCR was carried out using 35 cycles with 55 °C annealing temperature and 0.2 min extension time. The products were gel purified, phenol extracted, and used in a subsequent overlap PCR to generate longer fragments consisting of two adjacent small fragments. These fragments were cloned into a CDM7-derived plasmid containing a leader sequence of the CD5 surface molecule followed by a *Nhe*1/*Pst*1/*Mlu*1/*Eco*R1/*Bam*H1 polylinker. The correct sequence was confirmed by DNA sequencing. The gp120IIIb construct was generated by PCR using a *Sal*1/*Xho*1 HIV-1 HXB2 envelope fragment as template, followed by an exchange of a *Kpn*1/*Ear*1 fragment from a proviral clone. The wild-type gp120 mn constructs used as controls were cloned by PCR from HIV-1MN (NIH AIDS Repository) infected C8166 cells and contained either the native *env* leader or a CD5 leader sequence. Two clones of each construct were tested to avoid PCR-induced artefacts. The rat Thy-1 gene with the HIV envelope codon usage (rTHY-1env) was generated using three 150–170-mer oligonucleotides. In contrast to the syngp120mn, the PCR products were directly cloned and assembled in pUC12, and subsequently cloned into CDM7.

A GFP coding sequence was assembled in a similar manner from six fragments of approximately 120 bp each, using a strategy for assembly that relied on the ability of the restriction enzymes *Bsa*I and *Bbs*I to cleave outside their recognition sequence. Long oligonucleotides were synthesized which contained portions of the coding sequence of GFP embedded in flanking sequences encoding *Eco*RI and *Bsa*I sites at one end, and *Bam*HI and *Bbs*I sites at the other end, in the configuration *Eco*RI–*Bsa*I–GFP–*Bbs*I–*Bam*HI. The restriction site ends generated by the *Bsa*I and *Bbs*I sites were designed to yield compatible ends between adjacent GFP fragments, each of which was unique and non-selfcomplementary. The crude synthetic DNA segments were amplified by PCR, inserted between *Eco*RI and *Bam*HI in pUC9, and sequenced. Subsequently, the intact coding sequence was assembled in a six fragment ligation, using insert fragments prepared with *Bsa*I and *Bbs*I. Two of six plasmids resulting from the ligation bore an insert of correct size, and one contained the desired full length sequence. Mutation of Ser65 to Thr was accomplished by PCR, using a primer that overlapped a unique *Bss*SI site in the synthetic GFP.

Immunoprecipitation

293T cells were transfected by calcium phosphate using small dishes of 50–70 % confluent cells and 10 μg plasmid DNA. In cotransfection experiments with *rev*, cells were transfected with 10 μg gp120IIIb, gp120IIIbrre, syngp120mnrre or rTHY-1enveg1rre expression plasmid DNAs and 10 μg pCMVrev (AIDS Repository) or CDM7 plasmid DNA. After 48–60 h medium was exchanged and cells were incubated for additional 12 h in Cys/Met-free medium (Gibco) containing 200 μCi [^{35}S]Cys+Met (ICN). Supernatants were harvested and spun for 15 min at 3000 rpm to remove debris. After addition of the protease inhibitors leupeptin, aprotinin and PMSF (Sigma) to 2.5 μg ml^{-1}, 50 μg ml^{-1} and 100 μg ml^{-1}, respectively, 1 ml of supernatant was incubated with either 10 ml of packed protein A sepharose (Sigma) alone (rTHY-1enveg1rre) or with protein A sepharose and 3 mg of a purified CD4–immunoglobulin fusion protein (kindly provided by Behring) at 4 °C for 12 h. The protein A beads were washed 5 times with a buffer containing 100 mM Tris pH 7.5, 150 mM NaCl, 5 mM $CaCl_2$, 1 % NP-40. After the final wash, 10 μl of loading buffer containing 10 % glycerol, 4 % SDS, 4 % mercaptoethanol and 0.002 % bromphenol blue was added, samples were boiled for 3 min and applied on 7 % (all gp120 constructs) or 10 % (rTHY-1enveg1rre) SDS polyacrylamide gels. Gels were fixed in 10 % acetic acid and 10 % methanol, incubated with Amplify (Amersham) for 20 min, dried and exposed for 12 h.

ELISA

The concentration of gp120 in culture supernatants was determined using CD4-coated ELISA plates and goat anti-gp120 antisera in the soluble phase. Supernatants of 293T cells transfected by calcium phosphate were harvested after 4 days, spun at 3000 rpm for 10 min to remove debris and incubated for 12 h at 4° C on the plates (NEN Dupont). After 6 washes with PBS, 100 μl goat anti-gp120 antisera diluted 1:200 (AIDS Repository) were added for 2 h. The plates were washed again and incubated for 2 h with a peroxidase-conjugated rabbit anti-goat IgG antiserum 1:1000 (Cappel). Subsequently, the plates were washed and incubated for 30 min with 100 μl substrate solution containing 2 mg ml^{-1} o-phenylendiamine (Sigma) in sodium citrate buffer. The reaction was finally stopped with 100 μl 4 M sulfuric acid. Plates were read at 490 nm with a Coulter microplate reader. As a control, purified recombinant gp120IIIb (Repligen) was used.

Vaccinia virus recombinants

Recombinant vaccinia virus expressing syngp120mn under the control of the 7.5 promoter was generated by cloning a *Hin*d3/*Not*1 fragment containing the synthetic gp120mn gene into pTKG. Vaccinia recombination, plaque purification and generation of high-titer virus stocks was done in CV-1 cells essentially as described [64]. vPE-8 expresses HIV-1 IIIB gp120 under the 7.5 promoter [39]. Hela or 293 cells were infected at a multiplicity of infection of at least 10. Supernatants of ^{35}S-labelled cells were harvested 24 h post infection and immunoprecipitated as described above.

Immunofluorescence

293T cells were transfected by calcium phosphate and analyzed for surface Thy-1 expression after 3 days. Cells were detached with 1 mM EDTA in PBS and stained with the monoclonal antibody OX-7 (Accurate) and a FITC-conjugated goat anti-mouse immunoglobulin antiserum (Cappel). The analysis was done on a EPICS XL cytofluorometer (Coulter).

Compilation of codon frequency tables

Codon usage by retroviruses was compiled from the envelope precursor sequences of avian leukosis virus, bovine leukemia virus, feline leukemia virus, gibbon ape leukemia virus, human T-cell leukemia virus type I, Mason–Pfizer monkey virus, mouse mammary tumor virus, the 10A1 isolate of murine leukemia virus, the 4070A amphotropic isolate of MLV, rat leukemia virus, and simian sarcoma virus. The codon frequency tables for the non-HIV, non-SIV lentiviruses were compiled from the envelope precursor sequences for bovine immunodeficiency virus, caprine arthritis encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and Visna virus.

Acknowledgements

We thank Simon Wain-Hobson for a stimulating discussion and many suggestions, and Patricia Earl, Robert Gallo, Marie-Louise Hammarskjöld, Bernard Moss and David Rekosh for provision of reagents under the auspices of the NIH AIDS Research and Reference Reagent Program. J.H. was supported by a fellowship from the DKFZ. Portions of this work were supported by NIH grant HL53694 and an award to the Massachusetts General Hospital from Hoechst AG.

References

1. Daly TJ, Cook KS, Gray GS, Malone TE, Rusche JR: **Specific binding of HIV-1 recombinant rev protein to the Rev-responsive element *in vitro*.** *Nature* 1989, **342**:816–819.
2. Hadzopoulou-Cladaras M, Felber BK, Cladaras C, Athanassopoulos A, Tse A, Pavlakis GN: **The *rev* (trs/art) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a *cis*-acting sequence in the env region.** *J Virol* 1989, **63**:1265–1274.
3. Zapp ML, Green MR: **Sequence-specific RNA binding by the HIV-1 *rev* protein.** *Nature* 1989, **342**:714–716.
4. Malim MH, Hauber J, Le SY, Maizel JV, Cullen BR: **The HIV-1 *rev* transactivator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA.** *Nature* 1989, **338**:254–257.
5. Olsen HS, Nelbock P, Cochrane AW, Rosen CA: **Secondary structure is the major determinant for interaction of HIV *rev* protein with RNA.** *Science* 1990, **247**:845–848.
6. Huang XJ, Hope TJ, Bond BL, McDonald D, Grahl K, Parslow TG: **Minimal Rev-response element for type 1 human immunodeficiency virus.** *J Virol* 1991, **65**:2131–2134.
7. Kjems J, Brown M, Chang DD, Sharp PA: **Structural analysis of the interaction between the human immunodeficiency virus *rev* protein and the rev response element** *Proc Natl Acad Sci USA* 1991, **88**:683–687.
8. Feinberg MB, Jarrett RF, Aldovini A, Gallo RC, Wong- Staal F: **HTLV-III expression and production involve complex regulation at the levels of splicing and translation of viral RNA.** *Cell* 1986, **46**:807–817.
9. Sodroski J, Goh WC, Rosen C, Dayton A, Terwilliger E, Haseltine W: **A second post-transcriptional trans-activator gene required for HTLV-III replication.** *Nature* 1986, **321**:412–417.
10. Emerman M, Vazeux R, Peden K: **The rev gene product of the human immunodeficiency virus affects envelope-specific RNA localization.** *Cell* 1989, **57**:1155–1165.
11. Felber BK, Derse D, Athanassopoulos A, Campbell M, Pavlakis GN: **Cross-activation of the Rex proteins of HTLV- I and BLV and of the rev protein of HIV-1 and nonreciprocal interactions with their RNA responsive elements.** *New Biol* 1989, **1**:318–328.
12. Heaphy S, Dingwall C, Ernberg I, Gait MJ, Green,SM, Karn J, *et al*: **HIV-1 regulator of virion expression (Rev) protein binds to an RNA stem-loop structure located within the rev response element region.** *Cell* 1990, **60**:685–693.
13. Malim MH, Tiley LS, McCarn DF, Rusche JR, Hauber J, Cullen BR: **HIV-1 structural gene expression requires binding of the rev trans-activator to its RNA target sequence.** *Cell* 1990, **60**:675–683.
14. Malim MH, Cullen BR: ***rev* and the fate of pre- mRNA in the nucleus: implications for the regulation of RNA processing in eukaryotes** *Mol Cell Biol* 1993, **13**:6180–6189.
15. Fischer U, Meyer S, Teufel M, Heckel C, Luhrmann R, Rautmann G: **Evidence that HIV-1 rev directly promotes the nuclear export of unspliced RNA.** *EMBO J* 1994, **13**:4105–4112.
16. Bogerd HP, Fridell RA, Madore S, Cullen BR: **Identification of a novel cellular cofactor for the rev/rex class of retroviral regulatory proteins.** *Cell* 1995, **82**:485–494.
17. Arrigo SJ, Chen IS: ***rev* is necessary for translation but not cytoplasmic accumulation of HIV-1 vif, vpr, and env/vpu 2 RNAs.** *Genes Dev* 1991, **5**:808–819.
18. Lawrence JB, Cochrane AW, Johnson CV, Perkins A, Rosen CA: **The HIV-1 rev protein: a model system for coupled RNA transport and translation.** *New Biol* 1991, **3**:1220–1232.
19. D'Agostino DM, Felber BK, Harrison JE, Pavlakis GN: **The *rev* protein of human immunodeficiency virus type 1 promotes polysomal association and translation of gag/pol and vpu/env mRNAs.** *Mol Cell Biol* 1992, **12**:1375–1386.
20. Rosen CA: **HIV regulatory proteins: potential targets for therapeutic intervention.** *AIDS Res Hum Retroviruses* 1992, **8**:175–181.
21. Dayton AI, Terwilliger EF, Potz J, Kowalski M, Sodroski JG, Haseltine WA: **Cis-acting sequences responsive to the *rev* gene product of the human immunodeficiency virus J.** *AIDS* 1988, **1**:441–452.
22. Cochrane AW, Jones KS, Beidas S, Dillon PJ, Skalka AM, Rosen CA: **Identification and characterization of intragenic sequences which repress human immunodeficiency virus structural gene expression.** *J Virol* 1991, **65**:5305–5313.
23. Maldarelli F, Martin MA, Strebel K: **Identification of posttranscriptionally active inhibitory sequences in human immunodeficiency virus type 1 RNA: novel level of gene regulation.** *J Virol* 1991, **65**:5732–5743.
24. Schwartz S, Felber BK, Pavlakis GN: **Distinct RNA sequences in the gag region of human immunodeficiency virus type 1 decrease RNA stability and inhibit expression in the absence of *rev* protein.** *J Virol* 1992, **66**:150–159.
25. Brighty DW, Rosenberg M: **A *cis*-acting repressive sequence that overlaps the Rev-responsive element of human immunodeficiency virus type 1 regulates nuclear retention of env mRNAs independently of known splice signals.** *Proc Natl Acad Sci USA* 1994, **91**:8314–8318.
26. Zolotukhin AS, Valentin A, Pavlakis GN, Felber BK: **Continuous propagation of RRE(-) and Rev(-)RRE(-) human immunodeficiency virus type 1 molecular clones containing a cis- acting element of simian retrovirus type 1 in human peripheral blood lymphocytes.** *J Virol* 1994, **68**:7944–7952.
27. Barksdale SK, Baker CC: **The human immunodeficiency virus type 1 rev protein and the Rev-responsive element counteract the effect of an inhibitory 5′ splice site in a 3′ untranslated region.** *Mol Cell Biol* 1995, **15**:2962–2971.
28. Schwartz S, Campbell M, Nasioulas G, Harrison J, Felber BK, Pavlakis GN: **Mutational inactivation of an inhibitory sequence in human immunodeficiency virus type 1 results in Rev-independent gag expression.** *J Virol* 1992, **66**:7176–7182.
29. Olsen HS, Cochrane AW, Rosen C: **Interaction of cellular factors with intragenic cis-acting repressive sequences within the HIV genome.** *Virology* 1992, **191**:709–715.
30. Grantham R, Perrin P: **AIDS virus and HTLV-I differ in codon choices.** *Nature* 1986, **319**:727–728.
31. Kypr J, Mrazek J: **Unusual codon usage of HIV.** *Nature* 1987, **327**:20–20.
32. Sharp PM: **What can AIDS virus codon usage tell us?** *Nature* 1986, **324**:114–114.
33. Chou KC, Zhang CT: **Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implication.** *AIDS Res Hum Retrovir* 1992, **8**:1967–976.
34. Ward WW: **Energy transfer processes in bioluminescence.** *Photochem Photobiol Rev* 1979, **4**:1–57.
35. Prasher DC, Eckenrode VK, Ward WW, Prendergast FG, Cormier MJ: **Primary structure of the *Aequorea victoria* green-fluorescent protein.** *Gene* 1992, **111**:229–233.
36. Cody CW, Prasher DC, Westler WM, Prendergast FG, Ward WW: **Chemical structure of the hexapeptide chromophore of the *Aequorea* green-fluorescent protein.** *Biochemistry* 1993, **32**:1212–1218

37. Lasky LA, Groopman JE, Fennie CW, Benz PM, Capon DJ, Dowbenko DJ, *et al.*: **Neutralization of the AIDS retrovirus by antibodies to a recombinant envelope glycoprotein.** *Science* 1986, **233**:209–212.
38. Aruffo A, Stamenkovic I, Melnick M, Underhill CB, Seed B: **CD44 is the principal cell surface receptor for hyaluronate.** *Cell* 1990, **61**:1303–1313.
39. Earl PL, Koenig S, Moss B: **Biological and immunological properties of human immunodeficiency virus type 1 envelope glycoprotein: analysis of proteins with truncations and deletions expressed by recombinant vaccinia viruses.** *J Virol* 1991, **65**:31–41
40. Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC: **Green fluorescent protein as a marker for gene expression.** *Science* 1994, **263**:802–805.
41. Heim R, Cubitt AB, Tsien RY: **Improved green fluorescence.** *Nature* 1995, **373**:663–664.
42. Vartanian JP, Meyerhans A, Asjo B, Wain-Hobson S: **Selection, recombination, and G-A hypermutation of human immunodeficiency virus type 1 genomes.** *J Virol* 1991, **65**:1779–1788.
43. Gao F, Yue L, White AT, Pappas PG, Barchue J, Hanson AP, *et al.*: **Human infection by genetically diverse SIVSM-related HIV-2 in west Africa.** *Nature* 1992, **358**:495–499.
44. Vartanian JP, Meyerhans A, Sala M, Wain-Hobson,S: **G→A hypermutation of the human immunodeficiency virus type 1 genome: evidence for dCTP pool imbalance during reverse transcription.** *Proc Natl Acad Sci USA* 1994, **91**:3092–3096.
45. Wain-Hobson S, Sonigo P, Guyade M, Gazit A, Henry M: **Erratic G→A hypermutation within a complete caprine arthritis-encephalitis virus (CAEV) provirus.** *Virology* 1995, **209**:297–303.
46. Sala M, Wain-Hobson,S, Schaeffer F: **Human immunodeficiency virus type 1 reverse transcriptase tG:T mispair formation on RNA and DNA templates with mismatched primers: a kinetic and thermodynamic study.** *EMBO J* 1995, **14**:4622–4627.
47. Eden S, Cedar H: **Role of DNA methylation in the regulation of transcription.** *Curr Opin Genet Dev* 1994, **4**:255–259.
48. Holzschu DL, Martineau D, Fodor SK, Vogt VM, Bowser PR, Casey JW: **Nucleotide sequence and protein analysis of a complex piscine retrovirus, walleye dermal sarcoma virus.** *J Virology* 1995, **69**:5320–5331.
49. Blum HE, Harris JD, Ventura P, Walker D, Staskus K, Retzel E, Haase AT: **Synthesis in cell culture of the gapped linear duplex DNA of the slow virus visna.** *Virology* 1985, **142**:270–277.
50. Kuplec JJ, Tobaly-Taplero J, Canivet M, Santillana-Hayat M, Flugel RM, Peries J, Emanoil-Ravier R: **Evidence for a gapped linear duplex DNA intermediate in the replicative cycle of human and simian spumaviruses.** *Nucleic Acids Res* 1988, **16**:9557–9565.
51. Charneau P, Clavel F: **A single-stranded gap in human immunodeficiency virus unintegrated linear DNA defined by a central copy of the polypurine tract.** *J Virol* 1991, **65**:2415–2421.
52. Renshaw RW, Gonda MA, Casey JW: **Structure and transcriptional status of bovine syncytial virus in cytopathic infections.** *Gene* 1991, **105**:179–184.
53. Ikemura T: **Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs.** *J Mol Biol* 1982, **158**:573–597.
54. Grantham R, Gautier C, Gouy M: **Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type.** *Nucleic Acids Res* 1980, **8**:1893–1912.
55. Wain-Hobson S, Nussinov R, Brown RJ, Sussman, JL: **Preferential codon usage in genes.** *Gene* 1981, **13**:355–364.
56. Grosjean H, Fiers W: **Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes.** *Gene* 1982, **18**:199–209.
57. Ikemura T: **Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E coli* translational system.** *J Mol Biol* 1981, **151**:389–409.
58. Holm L: **Codon usage and gene expression.** *Nucleic Acids Res* 1986, **14**:3075–3087.
59. Makoff AJ, Oxer MD, Romanos MA, Fairweather NF, Ballantine S: **Expression of tetanus toxin fragment C in *E coli*: high level expression by removing rare codons.** *Nucleic Acids Res* 1989, **17**:10191–10202.
60. Williams DP, Regler D, Akiyoshi D, Genbauffe F, Murphy JR: **Design, synthesis and expression of a human interleukin-2 gene incorporating the codon usage bias found in highly expressed *Escherichia coli* genes.** *Nucleic Acids Res* 1988, **16**:10453–10467.
61. Ding D, Lipshitz HD: **Localized RNAs and their functions.** *Bioessays* 1993, **15**:651–658.
62. Kislauskis EH, Singer RH: **Determinants of mRNA localization.** *Curr Opin Cell Biol* 1992, **4**:975–978.
63. Narayan P, Ludwiczak RL, Goodwin EC, Rottman FM: **Context effects on N6-adenosine methylation sites in prolactin mRNA.** *Nucleic Acids Res* 1994, **22**:419–426.
64. Romeo C, Seed B: **Cellular immunity to HIV activated by CD4 fused to T cell or Fc receptor polypeptides.** *Cell* 1991, **64**:1037–1046.

Exhibit F

1: Mol Ther. 2000 Sep;2(3):288-97.

Fusion protein vectors to increase protein production and evaluate the immunogenicity of genetic vaccines.

Wu L, Barry MA.

Center for Cell and Gene Therapy, Baylor College of Medicine, Houston, Texas, 77030, USA.

Genetic immunization is a method for vaccination and laboratory antibody production where antigen-expressing plasmids are introduced into animals to elicit immune responses. Although genetic immunization works well for many antigens, problems can arise with protein sequences that (i) are toxic to host cells, (ii) are difficult to translate by mammalian cells, or (iii) evade immune presentation. We demonstrate here the ability to increase protein production and antigen secretion by the simple method of fusing poorly expressed sequences to well-expressed heterologous proteins. Proof-of-principle is demonstrated here using the poorly translated HIV-1 envelope whose protein production is rescued by fusing this antigen to the carboxy-termini of two well-expressed proteins: the cytoplasmic green fluorescent protein and the secreted human protein a1-antitrypsin. This approach represents a simple and substantially less expensive method to increase protein and antigen production than codon-optimization strategies. It may therefore be more useful than whole gene codon replacement to enable inexpensive laboratory antibody production of poorly expressed antigens and for large-scale genomic protein or antigen screening efforts. Finally, we demonstrate a second benefit of this antigen fusion strategy in which the test antigen is "sandwiched" between two positive control antigens. By this approach, we demonstrate the intrinsic lack of immunogenicity of HIV-1 envelope under conditions when robust antibody responses are generated against its fusion protein partners, but not against this evasive antigen. These fusion protein vectors therefore represent a simple approach to not only increase antigen production, but also assess antigen production and immunogenicity in vivo.

PMID: 10985959 [PubMed - indexed for MEDLINE]

Recent Activity

Turn Off | Clear

Fusion protein vectors to increase protein production and evaluate the immunogenicity of g...Fusion protein vectors to increase protein production and evaluate the immunogenicity of genetic vaccines.

High-level expression in mammalian cells of recombinant house dust mite allergen ProDer p ...High-level expression in mammalian cells of recombinant house dust mite allergen ProDer p 1 with optimized codon usage.

Codon optimization markedly improves doxycycline regulated gene expression in the mouse he...Codon optimization markedly improves doxycycline regulated gene expression in the mouse heart.

Gene optimization is necessary to express a bivalent anti-human anti-T cell immunotoxin in...Gene optimization is necessary to express a bivalent anti-human anti-T cell immunotoxin in Pichia pastoris.

Synonymous codon usage bias and the expression of human glucocerebrosidase in the methylot...Synonymous codon usage bias and the expression of human glucocerebrosidase in the methylotrophic yeast, Pichia pastoris.

» See more...

Display AbstractPlus Show 20 Sort By Send to

Proc. Natl. Acad. Sci. USA
Vol. 96, pp. 4482–4487, April 1999
Evolution

Exhibit G

Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*

LAURENT DURET AND DOMINIQUE MOUCHIROUD

Laboratoire de Biométrie, Génétique et Biologie des Populations, Unité Mixte de Recherche Centre National de la Recherche Scientifique 5558, Université Claude Bernard, 43 Bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

Edited by Samuel Karlin, Stanford University, Stanford, CA, and approved February 2, 1999 (received for review September 25, 1998)

ABSTRACT **We measured the expression pattern and analyzed codon usage in 8,133, 1,550, and 2,917 genes, respectively, from *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Arabidopsis thaliana*. In those three species, we observed a clear correlation between codon usage and gene expression levels and showed that this correlation is not due to a mutational bias. This provides direct evidence for selection on silent sites in those three distantly related multicellular eukaryotes. Surprisingly, there is a strong negative correlation between codon usage and protein length. This effect is not due to a smaller size of highly expressed proteins. Thus, for a same-expression pattern, the selective pressure on codon usage appears to be lower in genes encoding long rather than short proteins. This puzzling observation is not predicted by any of the current models of selection on codon usage and thus raises the question of how translation efficiency affects fitness in multicellular organisms.**

Nonrandom usage of synonymous codons is a widespread phenomenon, observed in genomes from many species in all domains of life. Such codon-usage biases may result from mutational biases, from natural selection acting on silent changes in DNA, or both. Selection on codon usage has been clearly demonstrated in several unicellular organisms (e.g., *Escherichia coli*, *Saccharomyces cerevisiae*) (for review, see ref. 1). Three characteristics of codon usage reflect such selective pressure in those organisms. First, codon usage is biased toward "preferred" codons that generally correspond to the most abundant tRNA species (2). Second, there is a positive correlation between codon-usage bias and the level of gene expression (3, 4). Finally, the rate of synonymous substitution between species is inversely correlated with codon usage bias, implying greater purifying selection on silent changes in highly biased genes (4, 5).

The selective differences between alternative synonymous codons are probably very small. Thus, in genes with low expression levels, or in species with small population sizes, selection is not sufficient to overcome genetic drift, and codon usage is essentially shaped by mutation patterns (for review, see ref. 6). The "selection–mutation–drift" model was proposed (4, 7, 8) to describe this balance between selection favoring optimal codons and mutation with drift allowing persistence of nonoptimal codons.

In multicellular eukaryotes, gene expression and tRNA abundance can be tissue- and developmental stage-specific and are difficult to quantify. However, the action of natural selection on codon usage has been established in *Drosophila melanogaster*: the limited data available show a relationship between codon preference and tRNA abundance (9, 10); negative correlations between codon usage bias and silent substitution rate have been observed in *Drosophila* (9, 11, 12); finally, anecdotal evidence suggests a relationship between codon-usage bias and gene expression level: genes known to be expressed at a high level, such as those encoding ribosomal proteins or glycolytic enzymes, show a greater-than-average codon bias (9, 12). Other studies, although less extensive, suggest that selection on codon usage may also occur in another invertebrate, *Caenorhabditis elegans* (13), and in the plant *Arabidopsis thaliana* (14).

Optimal codons probably confer fitness benefits by enhancing translation efficiency. However, it is not yet clear whether codon usage affects primarily the elongation rate, the cost of proofreading, or the accuracy of translation. Several studies suggested that in *D. melanogaster*, selection acts to increase translation accuracy (15, 16). However, in absence of expression data, it was not possible to directly test this hypothesis.

Recently, expressed sequence tag (EST) projects have been initiated in different species with the aim to make the inventory of all the mRNAs that they express. The thousands of obtained sequences are generally partial (typically, sequences are 300–500 nt long) and with a relatively high rate of sequencing errors (≈3%). However, these ESTs are accurate and long enough to unambiguously identify their corresponding genes. There is a high redundancy among those ESTs, which reflects the relative abundance of mRNAs in the tissue from which the cDNA library has been prepared. Thus, these data can be used to get rough estimates of gene expression patterns.

The purpose of the work presented here was to measure the expression levels of large sets of genes available for *D. melanogaster*, *C. elegans*, and *A. thaliana* to directly test whether there was selection on codon usage in those species. Our results demonstrate that selection acts on silent sites in those three distantly related multicellular eukaryotes. But surprisingly, we observed a strong negative correlation between codon-usage bias and protein length that is not due to a smaller size of highly expressed proteins. None of the current models of selection on silent site for translation efficiency accounts for this puzzling observation.

MATERIALS AND METHODS

Sequence Data. *C. elegans*, *D. melanogaster*, and *A. thaliana* sequences were extracted from GenBank release 105 (February 1998) (17), by using the ACNUC retrieval system (18). We selected complete protein-coding sequences (CDS) from nuclear genes, excluding pseudogenes and sequences described as ORFs or "unidentified reading frames". Histone genes also were excluded (see below). Only genomic sequences were selected, except for *D. melanogaster*, for which we also included CDS from mRNA sequences to increase the sample size. All CDS were compared with each other with BLASTN2 (19) to remove redundant sequences. In case of alternative splicing, we retained only the longest CDS variant. The final data set included 8,133, 1,550, and 2,917 CDS, respectively from *C. elegans*, *D. melanogaster*, and *A.*

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

PNAS is available online at www.pnas.org.

This paper was submitted directly (Track II) to the *Proceedings* office.
Abbreviations: CDS, protein-coding sequence; EST: expressed sequence tag; *Fav*, frequency of favored codons; RSCU, relative synonymous codon usage.
*To whom reprint requests should be addressed. e-mail: duret@biomserv.univ-lyon1.fr.

thaliana, among which 7,891, 439, and 2,386 were interrupted by introns.

Expression Profiles. Expression profiles were determined by counting the number of occurrence of each gene among EST sequences from different cDNA libraries that had been sampled with at least 9,000 ESTs. We selected in GenBank 67,987 *C. elegans* ESTs from whole-animal cDNA libraries at two developmental states (adult; embryo), 27,491 ESTs from *D. melanogaster* (adult ovary and head; embryo), and 36,207 ESTs from *A. thaliana* (one cDNA library pooled from four different tissues).

Selected CDS were first filtered with the XBLAST program (20) to mask repetitive elements. CDS were then compared with the species-specific EST data set by using BLASTN2 (19). BLASTN2 alignments showing at least 95% identity over 100 nt or more were counted as a sequence match. Because ESTs are derived from poly(A)$^+$ selected cDNA libraries, they cannot be used to estimate the abundance of replication-dependent histone mRNAs (that are not polyadenylated).

RESULTS

Measuring Gene Expression with ESTs. We selected from the databases 8,133, 1,550, and 2,917 complete protein-coding sequences from *C. elegans*, *D. melanogaster*, and *A. thaliana*, respectively. These large data sets represent 10–50% of the estimated number of genes in those three species, and are thus expected to be representative of whole genomes. EST sequences from different cDNA libraries were extracted from GenBank. For each species, and for each cDNA library, the expression level of selected genes was measured by counting the number of matching ESTs and dividing this number by the total number of ESTs sequenced in that cDNA library. Therefore, our measures reflect the relative mRNA abundance in those tissues where the cDNA libraries have been sampled. For *C. elegans* and *D. melanogaster*, libraries from two different stages (embryo, adult) were available. For genes expressed in both stages, we retained only their maximal relative abundance among these two cDNA libraries.

Genes were sorted according to their expression level and classified in four groups (Table 1). Genes without any EST made the first group. Expressed genes were classified in three other groups of low, moderate, and high expression. The limits between these classes were set for each species to obtain three samples of equal size (except for *A. thaliana*, where genes matching one single EST represent 53% of expressed genes).

It should be noted that estimates of expression level derived from ESTs are imprecise. Notably, 17% of *A. thaliana* ESTs were obtained from a normalized cDNA collection (i.e., from which redundant clones have been removed) (21). As a consequence, the mRNA abundance of genes expressed at a high level is underestimated in *A. thaliana*. However, the relative order of genes sorted according their expression level should not be affected. Thus, even if ESTs are partly normalized, the classification in four broad groups of expression that we used remains correct.

Frequency of Favored Codons and Gene Expression. Sharp and colleagues defined "optimal" codons as those showing a statistically significant increase in frequency between genes with low and high codon-usage bias (13, 22). Note that this definition does not necessarily imply any relationship with translation efficiency. In our opinion, the term optimal is equivocal here because it *a priori* suggests that there is a relationship with gene expression. Thus, codons found to occur significantly more often in highly than in lowly biased genes will hereafter be referred as favored codons. Here, optimal codons will refer only to those codons whose frequency has been shown to increase with gene expression.

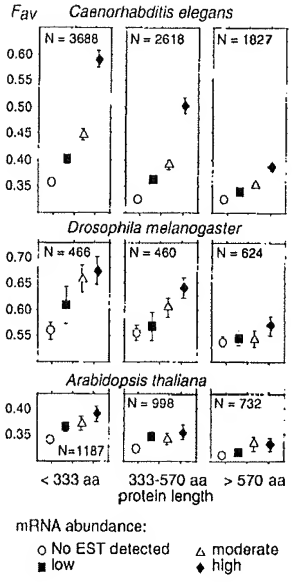


FIG. 1. Frequency of favored codons and gene expression in *C. elegans*, *D. melanogaster*, and *A. thaliana*. Average *Fav* values have been computed for different expression levels and protein lengths. Error bars indicate the 95% confidence interval.

Favored codons have been identified by multivariate analysis in *D. melanogaster* (22), *C. elegans* (13), and *A. thaliana* (14). We calculated for each gene the frequency of favored codons (*Fav*). *Fav* is a species-specific measure of codon-usage bias and is calculated as the number of occurrences of these favored codons divided by the total number of occurrences of the 18 amino acids having synonymous codons (13).

We computed average *Fav* in *C. elegans*, *D. melanogaster*, and *A. thaliana* for genes from the four groups of expression level. To investigate the effect of protein length on codon usage, the data set was split into three groups of equal sample size: short (<333 aa), intermediate, and long proteins (>570 aa) (Fig. 1).

Three observations can made. First, Fig. 1 clearly shows that in the three species studied, there is an increase of *Fav* with expression level. Thus, the relationship between codon-usage bias and gene expression that had been suggested for those multicellular organisms is here directly demonstrated. Second, the range of variation of *Fav* with expression level is not the same in all species: the increase is very sharp in *C. elegans* and *D. melanogaster* and less pronounced in *A. thaliana*. It should be noted however, that the weaker correlation between *Fav* and expression in this latter species may be partly because of the fact that mRNA abundance of highly expressed genes is underestimated (see above). Finally, the increase of codon-usage bias with expression level is much stronger in genes coding for short than for long proteins.

Detailed Analysis for Each Amino Acid. The relative synonymous codon usage (RSCU) is the observed frequency of a codon divided by the frequency expected if all synonyms for that amino acid were used equally. Thus, RSCU values close to 1.0 indicate

Table 1. Summary of relative mRNA abundance measures, based on EST sequence data

	Genes, no.					mRNA abundance × 10^5, expression			
			Expression						
Organism	All	No Est	Low	Moderate	High	All	Low	Moderate	High
C. elegans	8,133	4,495	1,150	1,242	1,246	9 (0–525)	4 (3–5)	10 (7–16)	44 (17–525)
D. melanogaster	1,550	675	214	330	331	16 (0–819)	7 (5–10)	13 (11–21)	57 (22–819)
A. thaliana	2,917	1,720	638	289	270	3 (0–146)	3 (3–3)	6 (6–6)	16 (8–146)

Number of genes and average relative mRNA abundance (range) are indicated for all genes and for each class of expression.

a lack of bias. The RSCU is independent of amino acid composition and is thus useful for comparing different sets of genes (13). Table 2 reports average RSCU values of all codons from *C. elegans*, *D. melanogaster*, and *A. thaliana* genes, according to their relative mRNA abundance and the length of the protein they code. ΔRSCU corresponds to the difference of RSCU values between genes expressed at a high level and those for which we did not detect any EST. Favored codons (identified by multivariate analysis) and optimal codons (the ones with positive ΔRSCU values) are highlighted. There is a remarkable correspondence between optimal and favored codons: in *C. elegans*, *D. melanogaster*, and *A. thaliana*, respectively, 100% (21/21), 91% (20/22), and 86% (18/21) of favored codons are also optimal, whereas only 13% (5/38), 14% (5/37), and 11% (4/38) of nonfavored codons are optimal (and in all of these latter cases, ΔRSCU is weak).

For all codons, ΔRSCU is strongest in *C. elegans* and weakest in *A. thaliana*, in agreement with the global measure of *Fav* (Fig. 1).

Generally, ΔRSCU values are higher in short than in long proteins. In these latter, many amino acids appear to have no optimal codons. Overall, long proteins do not have a particular set of optimal codons, but simply show a weakest codon-usage bias.

The usage of UAA terminator clearly increases with gene expression in *C. elegans* and *D. melanogaster* (Table 2). In *A. thaliana*, the terminator usage is less biased.

Mutational Bias or Selection? Biased codon usage may be explained either by selection on silent sites or by directional mutation pressure. The observed relationship between codon usage and mRNA abundance argues in favor of the selection model. However, it does not allow definitive rejection of the mutational bias hypothesis, because there may be a relationship between gene expression and mutation pattern. For example, it has been shown that the frequency of C-to-T mutations increases with the expression level of *E. coli* genes (23, 24).

If such expression-linked mutational bias was responsible for the biased nucleotide content at silent sites of highly expressed genes, it should affect not only exons but also introns. In the three species considered here, most optimal codons end in G or C, and thus G + C content at third codon position increases with expression level. But we did not detect any significant increase in introns G + C content with expression. On the contrary, intron G + C content decreases slightly (but significantly) with expression in *C. elegans* (data not shown). Therefore, the correlation between codon usage and expression is not due to a mutational bias, but to selection.

Codon Usage and Development Stages. We compared codon usage in *C. elegans* genes expressed only in embryo, only in adult or in both (respectively 898, 1,366, and 1,374 genes). For each of these subsets, we computed RSCU values for different gene expression levels and protein length. Optimal codons in embryo-specific and adult-specific genes are exactly the same ones as in genes expressed in both stages. Thus, there is no evidence of development stage-specific codon usage in *C. elegans*.

In multicellular organisms, selective pressure on codon usage is expected to depend not only on expression level of genes but also on the number of tissues or development stages where they are expressed. Indeed, we observed that *Fav* is in average much stronger in genes expressed both in embryo and adult than in stage-specific genes (Fig. 2). As noticed previously with data on mRNA abundance, the impact of expression pattern on codon usage is stronger in genes encoding short than long proteins.

Protein Length and Codon Usage. In all species, *Fav* decreases with the length of encoded proteins (Fig. 1). In genes of moderate or high expression (where the *Fav* variability is most pronounced), there is a significant negative correlation between Fav and the logarithm of protein length (Table 3).

We observed the same phenomenon for terminator usage. Among moderately or highly expressed genes, the optimal terminator (UAA) is used more frequently in genes encoding short than long proteins, both in *D. melanogaster* ($\chi^2 = 5.8$, $P = 0.016$) and *C. elegans* ($\chi^2 = 6.9$, $P = 0.009$).

In some species, codon usage has been shown to vary along the length of genes (25–27). In *D. melanogaster*, there seems to be an increase of G + C content at the start of genes followed by an overall decline (28). This decline affects not only exons, but also introns, and may be caused by a within-gene variation of mutational bias. Such effect could potentially be responsible for differences of G + C content between short and long genes. However, intron G + C content does not decrease with the length of the encoded protein (data not shown).

The observed decrease of the global *Fav* with protein length could be explained if the functional constraints responsible for selection against nonoptimal codons were restricted to a limited portion of the gene. To test this hypothesis, we measured the maximum *Fav* value (Fav_{max}) along coding sequences, by using a sliding window of 150 codons, moved by steps of three codons. Where the selective pressure on codon usage is weak (in *A. thaliana* and in genes expressed at a low level in *C. elegans* and *D. melanogaster*), we observed a slight increase of Fav_{max} with protein length (data not shown). Stochastic effects can probably explain this: the longer the sequence, the higher the chance of finding a segment of high frequency of favored codons. But in genes of moderate or high expression, Fav_{max} is significantly higher in short than in long proteins, both in *D. melanogaster* (Student's t test = 2.7, $P = 0.0067$) and *C. elegans* (Student's t test = 12.7, $P < 0.0001$). In *C. elegans* genes expressed at a high level, where the strongest effect was observed (Table 3), there is a significant negative correlation between Fav_{max} and protein length ($r = -0.42$, $P < 0.0001$).

Therefore, the negative correlation between *Fav* and protein length cannot be explained simply by localization of constraints on codon usage.

DISCUSSION

Selection for translation efficiency has been proposed for several years to explain codon usage bias in some multicellular eukaryotes (9, 13, 14). The large data set analyzed here shows a clear correlation between codon-usage bias and gene expression levels in *D. melanogaster*, *C. elegans*, and *A. thaliana* and thus provides direct evidence for selection on silent sites in those three distantly related organisms. It should be stressed that ESTs give only a rough picture of gene expression. Thus, we believe that in reality the correlation between codon-usage bias and expression may be even stronger than what we observed. In *C. elegans*, where expression data from adult and from embryo is available, we did not find any evidence of development stage-specific codon usage. However, the codon-usage bias is higher in genes expressed both in embryo and adult than in stage-specific genes. This is consistent with a selective pressure on codon usage depending not only on the expression level of genes but also on the number of tissues or development stages where they are expressed.

Surprisingly, we found that in the three species studied, the frequency of optimal codons decreases with the length of the encoded protein. A similar tendency has already been described in *D. melanogaster* and yeast (12, 16). But because the authors did not have expression data, they could not determine whether this correlation was a direct relationship between protein length and *Fav* or if it was caused by a tendency of genes expressed at a high level to encode short proteins. The authors retained this latter explanation and proposed that selection acts to reduce the length of proteins expressed at a high level (16). However, we did not find any evidence for such a selection (Table 4). In *A. thaliana*, there is no significant variation of protein length with expression level; in *D. melanogaster*, the only significant difference is that genes with no ESTs encode shorter proteins than expressed genes; and in *C. elegans*, there is an increase of average protein length with expression level. Indeed, for a same expression pattern, codon usage clearly decreases with increasing protein

Table 2. Codon usage (RSCU values), expression level, and protein length in *C. elegans*, *D. melanogaster*, and *A. thaliana* genes

	C. elegans					*D. melanogaster*					*A. thaliana*				
		Short proteins		Long proteins			Short proteins		Long proteins			Short proteins		Long proteins	
Codon	Favored	No EST	High	No EST	High	Favored	No EST	High	No EST	High	Favored	No EST	High	No EST	High
Arg		**25,872**	**3,833**	**17,254**	**40,260**		**2,404**	**1,042**	**12,349**	**8,030**		**8,004**	**1,372**	**20,572**	**2,478**
AGA		1.87	1.56	2.02	1.86		0.64	0.34	0.57	0.39		2.07	1.92	2.20	2.09
AGG		0.47	0.15	0.51	0.31		0.72	0.47	0.75	0.53	*	1.17^{+}	1.36	1.22	1.24
CGA		1.33	0.62	1.43	1.31		0.77	0.41	0.98	0.86		0.74	0.55	0.70	0.64
CGC	*	0.58^{+++}	1.34	0.45^{+}	0.57	*	1.92^{+++}	2.80	1.83^{+++}	2.29		0.43	0.41	0.40	0.37
CGG		0.51	0.24	0.51	0.39		0.83	0.47	0.97	0.79		0.53	0.37	0.57	0.54
CGU	*	1.24^{+++}	2.09	1.08^{+++}	1.56	*	1.11^{+++}	1.50	0.90^{++}	1.15	*	1.05^{+++}	1.40	0.91^{++}	1.11
Leu		**46,057**	**4,576**	**32,074**	**61,200**		**3,540**	**1,433**	**20,782**	**13,530**		**12,438**	**2,191**	**38,506**	**4,980**
CUA		0.58	0.19	0.63	0.43		0.46	0.23	0.56	0.51		0.62	0.52	0.63	0.59
CUC	*	0.96^{+++}	1.92	0.84^{++}	1.10	*	1.08	1.03	0.93	0.92	*	1.12^{+++}	1.50	0.92	0.96
CUG		0.76	0.69	0.79	0.73	*	2.64^{+++}	3.22	2.59	2.58		0.57	0.56	0.68	0.74
CUU	*	1.45^{+++}	1.76	1.41^{+++}	1.78		0.60	0.41	0.56	0.62		1.54^{+}	1.63	1.52	1.57
UUA		0.82	0.19	0.89	0.55		0.27	0.15	0.29	0.30		0.83	0.56	0.89	0.75
UUG		1.43	1.25	1.42	1.42		0.94	0.97	1.06	1.06		1.33	1.24	1.36	1.40
Ser		**40,824**	**4,049**	**28,694**	**58,229**		**3,295**	**1,121**	**21,409**	**11,905**		**12,722**	**2,316**	**35,838**	**4,297**
AGC		0.62^{+}	0.78	0.54	0.54		1.53	1.24	1.52	1.44	*	0.78	0.81	0.75	0.78
AGU		0.92	0.55	1.01	0.90		0.72	0.37	0.88	0.74		0.92	0.75	1.04	0.97
UCA		1.61	0.99	1.65	1.54		0.47	0.36	0.56	0.59		1.17	1.08	1.30	1.24
UCC	*	0.76^{+++}	1.39	0.73	0.73	*	1.76^{+}	1.92	1.37	1.43	*	0.78^{+}	0.89	0.67^{+}	0.76
UCG		0.79^{+}	0.98	0.76^{+}	0.87	*	1.03^{+++}	1.64	1.23	1.21		0.68	0.59	0.55	0.59
UCU		1.30	1.31	1.32^{+}	1.42		0.49	0.47	0.45^{+}	0.59		1.67^{++}	1.89	1.69	1.66
Thr		**30,391**	**3,260**	**20,939**	**41,866**		**2,498**	**918**	**13,725**	**8,234**		**7,401**	**1,452**	**19,911**	**2,690**
ACA		1.40	0.79	1.51	1.36		0.72	0.60	0.82	0.75		1.12	1.01	1.31	1.21
ACC	*	0.69^{+++}	1.50	0.59^{+}	0.67	*	1.74^{+++}	2.22	1.48^{+}	1.58	*	0.85^{+}	0.99	0.72^{+}	0.80
ACG		0.57	0.42	0.57	0.53		0.79	0.67	1.08	0.97		0.73	0.57	0.54	0.57
ACU		1.34	1.29	1.33^{+}	1.43		0.75	0.51	0.63	0.70		1.30^{+}	1.43	1.43	1.42
Pro		**23,312**	**2,961**	**15,553**	**34,714**		**2,442**	**809**	**13,932**	**7,782**		**6,716**	**1,373**	**18,010**	**2,424**
CCA	*	2.10^{+++}	2.82	2.03^{++}	2.31		0.93	0.78	1.01	0.97		1.27	1.25	1.35	1.38
CCC		0.38	0.19	0.39	0.29	*	1.52^{+++}	1.88	1.26	1.31	*	0.41^{+}	0.52	0.43	0.46
CCG		0.72	0.61	0.73	0.65		1.02	0.92	1.28	1.16		0.82	0.76	0.63	0.67
CCU		0.79	0.38	0.86	0.76		0.53	0.42	0.45^{+}	0.56		1.50	1.48	1.58	1.48
Ala		**29,935**	**5,149**	**19,697**	**49,193**		**3,432**	**1,494**	**18,160**	**11,932**		**8,775**	**2,171**	**23,348**	**3,535**
GCA		1.35	0.66	1.48	1.21		0.57	0.43	0.74	0.72		1.00	0.82	1.19	1.11
GCC	*	0.74^{+++}	1.40	0.66^{+}	0.74	*	2.01^{+++}	2.32	1.74	1.77	*	0.68^{+}	0.78	0.58	0.60
GCG		0.48	0.31	0.51	0.42		0.59	0.49	0.81	0.66		0.64	0.58	0.50	0.52
GCU	*	1.42^{++}	1.63	1.35^{++}	1.62		0.83	0.77	0.71^{+}	0.86		1.68^{+}	1.82	1.73	1.77
Gly		**25,766**	**4,491**	**16,959**	**39,020**		**3,281**	**1,284**	**15,131**	**10,315**		**9,656**	**2,068**	**23,343**	**3,579**
GGA	*	2.26^{+++}	2.77	2.28^{+++}	2.59		1.23	0.94	1.13	1.12		1.53	1.48	1.43	1.49
GGC		0.50	0.40	0.48	0.36	*	1.66^{++}	1.94	1.70	1.69	*	0.60	0.59	0.52	0.53
GGG		0.36	0.15	0.37	0.22		0.25	0.16	0.32	0.23		0.54	0.47	0.69	0.61
GGU		0.87	0.69	0.87	0.83		0.86^{+}	0.96	0.84^{+}	0.96	*	1.34^{+}	1.46	1.36	1.37
Val		**31,831**	**3,934**	**21,386**	**45,812**		**2,662**	**1,123**	**13,116**	**8,981**		**9,361**	**1,784**	**25,453**	**3,497**
GUA		0.67	0.25	0.77	0.59		0.30	0.28	0.43	0.44		0.56	0.47	0.67	0.60
GUC	*	0.83^{+++}	1.56	0.74^{+}	0.88	*	1.13^{+}	1.29	0.93	0.94	*	0.81	0.87	0.69	0.69
GUG		0.90	0.71	0.88	0.79	*	1.79	1.84	1.92	1.81		1.07	1.11	1.02	1.06
GUU		1.60	1.47	1.61^{+}	1.74		0.78	0.58	0.72^{+}	0.80		1.56	1.56	1.63	1.64
Lys		**34,361**	**4,512**	**22,049**	**47,882**		**2,529**	**1,441**	**11,653**	**9,450**		**9,257**	**1,930**	**24,600**	**3,135**
AAA		1.27	0.59	1.31	1.11		0.59	0.35	0.60	0.54		0.96	0.80	1.00	0.94
AAG	*	0.73^{+++}	1.41	0.69^{++}	0.89	*	1.41^{++}	1.65	1.40	1.46	*	1.04^{+}	1.20	1.00	1.06
Asn		**25,983**	**2,605**	**18,131**	**34,502**		**1,936**	**718**	**11,562**	**7,439**		**6,174**	**998**	**17,808**	**2,166**
AAC	*	0.75^{+++}	1.22	0.70	0.76	*	1.23^{++}	1.47	1.09^{+}	1.17	*	1.03^{+}	1.15	0.90^{+}	0.99
AAU		1.25	0.78	1.30	1.24		0.77	0.53	0.91	0.83		0.97	0.85	1.10	1.01
Gln		**19,766**	**2,424**	**13,462**	**32,392**		**2,071**	**665**	**14,445**	**8,767**		**4,845**	**875**	**14,059**	**1,783**
CAA		1.37	1.19	1.39	1.37		0.55	0.41	0.60	0.58		1.18	1.07	1.12	1.07
CAG	*	0.63^{+}	0.81	0.61	0.63	*	1.45^{+}	1.59	1.40	1.42	*	0.82^{+}	0.93	0.88	0.93
His		**11,626**	**1,270**	**7,872**	**17,198**		**1,136**	**342**	**7,300**	**3,789**		**3,222**	**606**	**8,933**	**1,100**
CAC	*	0.75^{+++}	1.23	0.72	0.73	*	1.25^{++}	1.46	1.21	1.22	*	0.81^{+}	0.90	0.69^{+}	0.82
CAU		1.25	0.77	1.28	1.27		0.75	0.54	0.79	0.78		1.19	1.10	1.31	1.18
Glu		**30,622**	**3,979**	**21,650**	**55,173**		**2,508**	**1,231**	**13,586**	**11,132**		**9,708**	**1,623**	**26,319**	**3,529**
GAA		1.30	0.92	1.36	1.30		0.58	0.35	0.62	0.61		1.02	1.00	1.08	1.00
GAG	*	0.70^{+++}	1.08	0.64	0.70	*	1.42^{++}	1.65	1.38	1.39	*	0.98	1.00	0.92^{+}	1.00
Asp		**25,098**	**3,551**	**17,514**	**43,357**		**2,166**	**1,052**	**11,557**	**8,516**		**7,676**	**1,366**	**20,700**	**2,743**
GAC	*	0.66^{++}	0.92	0.60	0.58	*	0.98^{+}	1.08	0.95	0.96	*	0.64^{+}	0.72	0.60	0.63
GAU		1.34	1.08	1.40	1.42		1.02	0.92	1.05	1.04		1.36	1.28	1.40	1.37
Tyr		**18,124**	**1,765**	**11,538**	**20,250**		**1,407**	**500**	**6,456**	**3,871**		**4,061**	**774**	**11,149**	**1,403**
UAC	*	0.84^{+++}	1.35	0.80^{+}	0.89	*	1.35^{++}	1.57	1.24	1.29	*	0.97^{++}	1.19	0.86	0.93
UAU		1.16	0.65	1.20	1.11		0.65	0.43	0.76	0.71		1.03	0.81	1.14	1.07
Cys		**12,135**	**1,016**	**7,540**	**12,899**		**1,174**	**244**	**4,433**	**2,242**		**2,768**	**439**	**7,481**	**755**
UGC	*	0.87^{+++}	1.28	0.82	0.84	*	1.44^{++}	1.68	1.44	1.42	*	0.80^{+}	0.89	0.77^{+}	0.86
UGU		1.13	0.72	1.18	1.16		0.56	0.32	0.56	0.58		1.20	1.11	1.23	1.14
Phe		**29,190**	**2,470**	**17,908**	**28,498**		**1,536**	**599**	**7,614**	**4,898**		**6,274**	**1,039**	**16,537**	**2,103**
UUC	*	0.90^{+++}	1.50	0.88^{++}	1.17	*	1.34^{++}	1.59	1.27	1.31	*	0.98^{++}	1.18	0.87	0.93
UUU		1.10	0.50	1.12	0.83		0.66	0.41	0.73	0.69		1.02	0.82	1.13	1.07
Ile		**34,721**	**3,057**	**23,018**	**40,126**		**2,136**	**937**	**10,904**	**7,260**		**7,243**	**1,288**	**20,984**	**2,753**
AUA		0.54	0.10	0.57	0.32		0.51	0.22	0.60	0.49		0.72	0.50	0.79	0.62
AUC	*	0.84^{+++}	1.75	0.80^{++}	1.01	*	1.59^{+++}	1.95	1.41	1.48	*	1.09^{++}	1.33	0.94^{+}	1.08
AUU		1.62	1.16	1.63	1.66		0.90	0.83	0.99	1.03		1.20	1.17	1.28	1.29

Table 2. (*Continued*)

		C. elegans					*D. melanogaster*					*A. thaliana*			
		Short proteins		Long proteins			Short proteins		Long proteins			Short proteins		Long proteins	
Codon	Favored	No EST	High	No EST	High	Favored	No EST	High	No EST	High	Favored	No EST	High	No EST	High
Ter		**2,543**	**298**	**429**	**663**		**255**	**80**	**232**	**140**		**680**	**122**	**458**	**59**
UAA		1.43^{+++}	2.00	1.37^{+++}	1.75		1.53^{+++}	1.95	1.14^{+++}	1.59		1.29^{+}	1.40	1.19	0.86
UAG		0.51	0.53	0.58	0.52		0.73	0.79	1.00	0.69		0.64	0.39	0.50^{+}	0.66
UGA		1.06	0.46	1.05	0.73		0.74	0.26	0.87	0.73		1.08^{+}	1.20	1.31^{+}	1.47

No EST and high denote genes with very low (no detected EST) or high expression level. Short and Long proteins denote genes with, respectively, <333 and >570 codons. Favored codons (defined by multivariate analysis) are indicated * (data from Refs. 13, 14, and 21). Optimal codons, defined by a higher RSCU value in highly expressed genes, are indicated +++ (ΔRSCU ≥ 0.30), ++ (ΔRSCU ≥ 0.20), and + (ΔRSCU ≥ 0.08). Numbers in boldface represent the number of codons analyzed.

length (Fig. 1, 2). Thus, the selective pressure on codon usage is lower in genes encoding long than short proteins.

It is likely that in the three species studied here, selection on codon usage acts to increase translation efficiency, as in other organisms such as yeast or *E. coli* (2–4). It is thought that the use of codons that match the most abundant tRNA reduces the time to find and bind the correct tRNA. Hence, optimal codons not only increase the elongation rate but also decrease the likelihood of binding a noncognate tRNA. Thus, different models have been proposed to account for the effect of translation efficiency on fitness: selection may act to maximize elongation rate, minimize the cost of proofreading, or maximize the accuracy of translation (8). All three models predict that codon-usage bias should be higher in genes expressed at high levels (8, 15, 29). But interestingly, none of these models accounts for the protein length effect on codon usage observed here. These different models are discussed below.

Powell and Moriyama (12) hypothesized that this length effect could be explained by selection for translation rate. Assume for example that a nonoptimal codon requires twice as long to incorporate an amino acid as does the optimal codon. Mutations to nonoptimal codons will have a greater relative effect in smaller genes than in longer ones: in a short gene with 100 codons, such mutation would increase translation time by 1%, whereas the same mutation in a gene with 1,000 codons would increase translation time by only 0.1%. Thus, such mutations are more likely to be counterselected in short genes than in long ones.

However, several arguments suggest that this model is not realistic. First, it is unlikely that selection acts to increase the rate of synthesis of a particular protein (except in specialized tissues that are devoted to the production of a few extremely abundant proteins such as silk glands in *Bombyx mori*) (30). Rather, selection is thought to act to increase the cell growth rate. Therefore, selection should act to increase the production of all cell constituents, and not of a particular protein species. Moreover, it seems that initiation is the limiting step in protein translation (reviewed in refs. 1 and 8). Thus, the rate of production of a given protein is determined by the initiation rate and not by the elongation rate. Indeed, the effect of codon usage on protein synthesis is thought to be indirect: the use of optimal codons in a given gene will increase the elongation rate and thus reduce the time the ribosome is bound to the mRNA: this leads to an increase in the pool of free ribosomes and hence to an increase in the translation initiation rate of all mRNA species (1, 8). Hence, the use of optimal codon in a given gene is thought to increase the production of all proteins in a cell, not only its own protein product. The delay caused by nonoptimal codons—and the resulting decrease in concentration of free ribosomes—will be the same in long and short mRNAs. Thus, the strength of selection on codon usage for an increased rate of protein production should be independent of protein length.

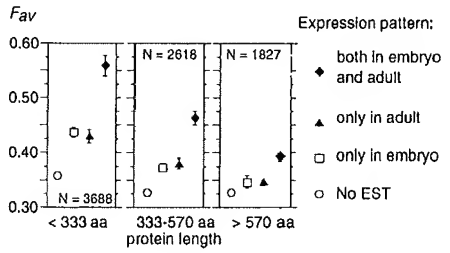


FIG. 2. Frequency of favored codons and development stage expression in *C. elegans*. Average *Fav* values have been computed for different patterns of expression and protein lengths. Error bars indicate the 95% confidence interval.

Selection to minimize the cost of proofreading is also expected to be independent of protein length. The process of rejecting noncognate tRNAs decreases elongation rate and consumes energy. The energy waste caused by a nonoptimal codon should be the same in a mRNA encoding a long or a short protein. And, as explained above, the strength of selection on elongation rate does not depend on protein length.

The model of selection on translation accuracy predicts that codon-usage bias should be higher in long than in short proteins: because the cost of producing a protein is proportional to its length, selection in favor of codons that increase accuracy should be higher in longer genes, as has been observed in *E. coli* (16, 29). This model is thus in total contradiction with our observations. Therefore, even if there is evidence for selection on translation accuracy in *D. melanogaster* (15), this is not the major factor that shapes codon usage in the three organisms studied here.

Another possible explanation for this length effect comes from population genetics theory. In a simulation study, Li (7) noticed that if the selective advantages conferred by optimal codons are additive (which is the case for the three models mentioned above) and if all sites within a gene are completely linked (i.e., recombination within a gene is rare relatively to mutation), then the efficacy of selection on optimal codons decreases with increasing protein length. This interference between linked sites is analogous to Muller's ratchet effect (31, 32): the efficacy of natural selection is reduced when genetic linkage exists among multiple sites affected by selection. A prediction of that model is that two tightly linked genes affected by selection on codon usage (i.e., highly expressed) should also interfere and behave as if they were a single gene encoding a longer protein. Thus, a long gene having a short neighbor should have the same codon usage as a short gene having a long neighbor. In the *C. elegans* data set, 449 genes expressed at high level have a neighbor that is expressed at a high level, <5 kb from their 5′ or 3′ end. Fig. 3 clearly shows that their codon usage is affected by their own length but not by the length of their neighbor. Therefore, neighbor genes do not seem to interfere. It is unlikely that recombination is frequent enough so

Table 3. Linear correlations between *Fav* and protein length (logarithmic scale)

	Expression	
Organism	Moderate	High
C. elegans	$R = -0.45$	$R = -0.60$
D. melanogaster	$R = -0.48$	$R = -0.41$
A. thaliana	$R = -0.33$	$R = -0.37$

$P < 0.0001$ for all values.

Table 4. Protein length and gene expression

Organism	Protein length, amino acids				
	All genes	Without EST	Expression		
			Low	Moderate	High
C. elegans	445 ± 388	336 ± 221	452 ± 369	542 ± 408	734 ± 617
D. melanogaster	624 ± 546	541 ± 481	731 ± 591	686 ± 583	662 ± 577
A. thaliana	452 ± 311	459 ± 316	456 ± 322	429 ± 273	425 ± 285

Values shown are mean ± SD.

that two genes within 5 kb are not genetically linked. If this were the case, we would expect that recombination should also occur within genes (that are 2.4 kb long on average), and hence the Li effect should not be detected. Moreover, even when considering closer genes, we do not find any evidence of interference: genes expressed at high level that have a very close neighbor with genes expressed at high level (<2 kb, $n = 229$) do not have a lower codon usage than genes expressed at high level that have no neighbor with genes expressed at a high or moderate level within 5 kb ($n = 409$) (t test = 0.4 $P = 0.69$).

In conclusion, none of the current models of selection on codon usage is consistent with the observed decrease of codon-usage bias in genes encoding long proteins. This length effect has been observed both in plant and metazoan species. It seems to occur also in yeast but not in *E. coli*, where codon-usage bias increases with protein length (16, 29). Thus, the selective pressure acting on codon usage may be different in eukaryotes and eubacteria.

The finding of selection on codon usage in very distantly related taxa, such as plants and animals, suggests that this is a widespread phenomenon. However, we did not observe any correlation between codon usage and expression level in human genes (33) (unpublished data). As noticed by others (6, 9), this absence of selection may be explained by population genetics: a mutation that is advantageous in a species with large effective population size may be neutral in a small population, where random drift overcomes selection. In mammals, effective population sizes have been estimated to be $\approx 10^4$, i.e., 10^2 to 10^3 smaller than in *Drosophila* species (34). Therefore, it is likely that in most human genes, fitness differences among synonymous codons is not sufficient to overcome drift.

A prediction of our observation is that the genes on which selection for optimal codons could operate in mammals should be short and expressed at very high levels in many different tissues and development stages. Indeed, to our knowledge, the only example of selection on silent site in mammals was described in H3 histones, that are short genes (137 codons) and expressed at extremely high level during S phase of the cell cycle in every cell

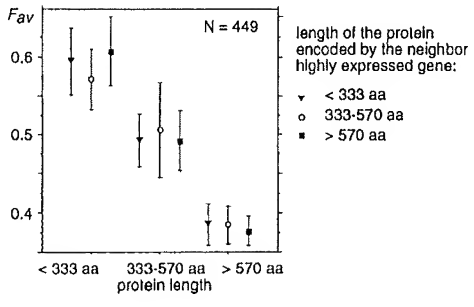


FIG. 3. Frequency of favored codons in *C. elegans* genes expressed at high level having a neighbor expressed at high level less than 5 kb from their 5′ or 3′ end. Average *Fav* values have been computed for different protein lengths. Error bars indicate the 95% confidence interval.

of the animal (35). How translation efficiency affects fitness in eukaryotes, however, remains unexplained.

We are grateful to M. Gouy, C. Gautier, and A. Eyre-Walker for many helpful discussions. This work is supported by the CNRS (Centre National de la Recherche Scientifique).

1. Kurland, C. G. (1991) *FEBS Lett.* **285,** 165–169.
2. Ikemura, T. (1992) in *Transfer RNA in Protein Synthesis*, eds. Hatfield, D. L., Lee, B. J. & Pirtle, R. M. (CRC, Boca Raton, FL), pp. 87–111.
3. Gouy, M. & Gautier, C. (1982) *Nucleic Acids Res.* **10,** 7055–7074.
4. Sharp, P. M. & Li, W. H. (1986) *J. Mol. Evol.* **24,** 28–38.
5. Sharp, P. M. & Li, W. H. (1987) *Mol. Biol. Evol.* **4,** 222–230.
6. Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. & Peden, J. F. (1995) *Philos. Trans. R. Soc. London B* **349,** 241–247.
7. Li, W. H. (1987) *J. Mol. Evol.* **24,** 337–345.
8. Bulmer, M. (1991) *Genetics* **129,** 897–907.
9. Shields, D. C., Sharp, P. M., Higgins, D. G. & Wright, F. (1988) *Mol. Biol. Evol.* **5,** 704–716.
10. Moriyama, E. N. & Powell, J. R. (1997) *J. Mol. Evol.* **45,** 514–523.
11. Sharp, P. M. & Li, W. H. (1989) *J. Mol. Evol.* **28,** 398–402.
12. Powell, J. R. & Moriyama, E. N. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 7784–7790.
13. Stenico, M., Lloyd, A. T. & Sharp, P. M. (1994) *Nucleic Acids Res.* **22,** 2437–2446.
14. Chiapello, H., Lisacek, F., Caboche, M. & Henaut, A. (1998) *Gene* **209,** GC1–GC38.
15. Akashi, H. (1994) *Genetics* **136,** 927–935.
16. Moriyama, E. N. & Powell, J. R. (1998) *Nucleic Acids Res.* **26,** 3188–3193.
17. Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J. & Ouellette, B. F. F. (1998) *Nucleic Acids Res.* **26,** 1–7.
18. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. & Di-Paola, G. (1985) *Comp. Appl. Biosci.* **1,** 167–172.
19. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
20. Claverie, J.-M. & States, D. J. (1993) *Comput. Chem.* **17,** 191–201.
21. Cooke, R., Raynal, M., Laudie, M., Grellet, F., Delseny, M., Morris, P. C., Guerrier, D., Giraudat, J., Quigley, F., Clabault, G., *et al.* (1996) *Plant J.* **9,** 101–124.
22. Sharp, P. M. & Lloyd, A. T. (1993) in *An Atlas of Drosophila Genes: Sequences and Molecular Features*, ed. Maroni, G. (Oxford Univ. Press, New York), pp. 378–397.
23. Beletskii, A. & Bhagwat, A. S. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 13919–13924.
24. Beletskii, A. & Bhagwat, A. S. (1998) *Biol. Chem.* **379,** 549–551.
25. Bulmer, M. (1988) *J. Theor. Biol.* **133,** 67–71.
26. Chen, G. F. & Inouye, M. (1990) *Nucleic Acids Res.* **18,** 1465–1473.
27. Eyre-Walker, A. & Bulmer, M. (1993) *Nucleic Acids Res.* **21,** 4599–4603.
28. Kliman, R. M. & Eyre-Walker, A. (1998) *J. Mol. Evol.* **46,** 534–541.
29. Eyre-Walker, A. (1996) *Mol. Biol. Evol.* **13,** 864–872.
30. Garel, J. P., Hentzen, D. & Daillie, J. (1974) *FEBS Lett.* **39,** 359–363.
31. Muller, H. J. (1964) *Mutat. Res.* **1,** 2–9.
32. Felsenstein, J. (1974) *Genetics* **78,** 737–756.
33. Karlin, S. & Mrázek, J. (1996) *J. Mol. Biol.* **262,** 459–472.
34. Nei, M. & Graur, D. (1984) *Evol. Biol.* **17,** 73–118.
35. Debry, R. W. & Marzluff, W. F. (1994) *Genetics* **138,** 191–202.